

UNIVERSITÉ DE PARIS VIII - VINCENNES-SAINT-DENIS

U.F.R. Langage Informatique Technologie

N° attribué par la bibliothèque

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS VIII

Discipline : Sciences de l'information et de la communication

Présentée et soutenue publiquement

par

Jean BAUDOIN (Waris Abdukerim JANBAZ, Waresijiang ABUDUKELIMU)

Octobre 2009

Titre :

Recherches sur le traitement informatique d'une langue turcique agglutinante : l'ouïghour

Directeur de recherche :

Imad Saleh, Université Paris VIII

Jury :

Mohamed Hassoun, Prof., ENSSIB

Said Tazi, MCF/HDR, Université des Sciences Sociales Toulouse 1

Khalidoun Zreik, Prof., Université de Paris VIII

Mohamed Quafafou, Prof., Université de la Méditerranée

Jean Rahman Duval, Dr., Expert linguistique OCDE

Imad Saleh, Prof, Dir., Université de Paris VIII

À mon people et ma langue menacés ...

Remerciements

Ma plus grande gratitude va à Monsieur Imad Saleh, directeur du Département Hypermédia de l'Université Paris VIII, qui m'a accordé sa confiance en acceptant d'être mon directeur de thèse et m'a accueilli au sein du laboratoire Paragraphe, pour son soutien constant, son attention, sa disponibilité et ses précieux conseils. Ce travail n'aurait pu être mené à son terme sans ses nombreux encouragements. Je le remercie profondément et de tout mon cœur.

Je tiens à remercier monsieur Jean Rahman Duval, qui m'a beaucoup encouragé et aidé dans la poursuite de mes études universitaires. Il ne m'a jamais dit non, même lorsque je lui ai envoyé des textes à corriger à la dernière minute. Je ne pourrai jamais assez le remercier.

Mes remerciements s'adressent aussi aux Professeurs Khaldoun Zreik, Mohamed Quafafou et Docteur Jean Rahman Duval pour leur participation en tant que membres de jury et aux Professeurs Mohamed Hassoun et Said Tazi pour avoir accepté d'être rapporteur de cette thèse.

Tous mes remerciements vont aussi à l'équipe Hypertexte Dynamique du Laboratoire Paragraphe qui m'a accueilli en son sein.

J'aimerais remercier également messieurs Scott Wade, Bernhard Rieder, Nasreddine Bouhaï et Everardo Reyes García pour leur amitié et leurs encouragements.

J'aimerais exprimer mon affectueuse gratitude à ma mère, Gülnisa Musa, et à toute la famille Janbaz pour son soutien ininterrompu, malgré les milliers de kilomètres qui nous séparent.

Enfin, un grand merci à Dilnur Kahar, mon épouse qui a donné naissance à mon fils, Oyghan, pour son soutien et sa patience pendant les moments dramatiques de la guerre de nerfs que peut devenir la rédaction d'une thèse.

Résumé

Cette étude présente les caractéristiques, les écritures et la structure de la langue ouïghoure en faisant une étude linguistique et en proposant de nouveaux modèles expérimentaux qui faciliteront le développement des outils informatiques et le traitement automatique de la langue afin de contribuer à l'informatisation de la langue ouïghoure.

Plus précisément, notre étude consiste en quatre parties : la première partie présente les problématiques d'étude, les caractéristiques de la langue et des écritures, notamment le processus d'unification de l'écriture ouïghoure-latine ; la deuxième partie expose les notions de base d'extraction d'information et démontre la possibilité d'extraction d'entités nommées en utilisant un outil d'extraction, afin d'expérimenter les conceptions et les théories proposées ; la troisième partie est consacré à l'étude linguistique notamment sur l'aspect agglutinant de la langue et les règles morphologiques de suffixation qui seront appliquées pendant la réalisation des outils prototypes proposés dans cette thèse ; enfin la quatrième partie mettre en évidence les problématiques de traitement de la langue ouïghoure dans une situation où les systèmes d'exploitation ne supporte pas la langue ouïghoure. Dans cette partie, nous décrivons les difficultés existantes et nous proposerons des solutions innovantes afin de les résoudre dans les domaines suivants :

- Unification des polices et création d'une police ouïghoure basée sur l'Unicode
- Implémentation des méthodes d'entrées au niveau système et au niveau navigateur
- Création des convertisseurs multiécriture
- Réalisation d'un dictionnaire ouïghour – anglais en ligne
- Mise en place d'un générateur lexical basé sur les règles morphologiques de suffixation de l'ouïghour
- Développement d'un analyseur et explorateur de suffixes
- Démonstration d'extraction de l'information
- Implémentation d'un parseur et un correcteur d'orthographe

Mots clés : Langue agglutinante, suffixe, analyseur, écriture, Unicode, police, encodage, méthode d'entrée, extraction, parseur, correcteur d'orthographe, dictionnaire en ligne, convertisseur

Abstract

This study presents the characteristics, writing systems and structure of Uyghur language by doing a linguistic study. Our approach will consist of new trial models that facilitate the development and realization of Uyghur software tools, and contribute to the Uyghur information technology.

More precisely, our study consists of four phases: Firstly, we are going to present the main issues of the study, characteristics of the language and its writing systems, especially the unification procedure of the Latin-Script Uyghur. Secondly, we briefly introduce some basic notions for the retrieval of information, and we will do a demonstration of named entities retrieval, using an extraction tool, in order to test concepts and theories that we are proposing. Then, we will discuss linguistic issues – mainly on the agglutinative aspect and morphological suffixation rules – which are applied during the implementation of prototype tools proposed in this study. Finally, we underline problems in natural language processing (NLP) created by Uyghur language and non-Uyghur supporting environments. We will discuss the existing difficulties and we will suggest innovative solutions to resolve such problems with the following fields:

- Standardization of Uyghur fonts and creation of a Unicode based Uyghur font
- Implementation of system-level and browser-level input methods and
- Creation of multi-script converting tools
- Realization of an online Uyghur – English dictionary
- Implementation of a lexical generator based on the morphological suffixation rules of Uyghur
- Design and creation of a suffix analyzer and an explorer
- Demonstration of Uyghur information retrieval
- Implementation of a parser and spell checker

Keywords: Agglutinative language, suffix, analyzer, writing system, Unicode, font, encoding, input method, extraction, parser, spell checker, online dictionary, converter

REMERCIEMENTS.....	3
RESUME.....	4
ABSTRACT.....	5
CHAPITRE 1. INTRODUCTION	10
1.1 Les révolutions de l’informatisation ouïghoure	10
1.1.1 Qu’est que l’ouïghour ?.....	10
1.1.2 Informatisation et langue ouïghoure	11
1.2 Bref état des recherches	13
1.3 Problématique de la recherche.....	14
1.4 Plan de thèse.....	15
CHAPITRE 2. LA GESTION DU SAVOIR.....	17
2.1 Définition et enjeux.....	17
2.1.1 La recherche	17
2.1.2 L’information	17
2.1.3 Définition d’extraction d’information	19
2.1.4 Définition d’un Système de Recherche d’Information	20
2.1.5 Rappel et Précision	21
2.1.6 Indexation.....	22
2.2 Particularités de l’ouïghour	23
2.2.1 Écritures ouïghoures.....	23
2.2.2 Caractéristiques techniques de l’écriture.....	28
2.2.2.1 Nombre de signes	28
2.2.2.2 Type d’écriture.....	29
2.2.2.3 Direction de l’écriture	29
2.2.2.4 Caractères de ponctuations et des chiffres.....	29
2.2.2.5 La spécificité de l’écriture ouïghoure	29
2.2.2.6 Identification de la langue ouïghour	30
2.2.2.7 Méthodes d’entrée de l’ouïghour	30
2.2.2.8 Stockage et bases de données	31
2.2.3 Extension territoriale et langues notées	31
2.3 La spécificité de l’ouïghour.....	32
2.4 Un regard sur les langues turques	33
CHAPITRE 3. UNE INTRODUCTION A L’ECRITURE OUÏGHOURE-LATINE	38
3.1 Introduction	38
3.2 Historique	40
3.3 Principes de base utilisés dans la création du LSU	44

3.4	Lettres ne posant pas de problème particulier.....	47
3.5	Lettres problématiques et décisions	47
3.6	Règles d'orthographe du LSU	53
3.7	Statut actuel et perspectives futures.....	55

CHAPITRE 4. TRAITEMENT DE LA LANGUE OUÏGHOURE..... 57

4.1	Introduction	57
4.2	Historique	57
4.4	Autres sources de problèmes.....	63
4.5	Développement des polices ouïghoures et une méthode d'entrée au niveau du système d'exploitation	67
4.6	Incorporation des polices et l'affichage des lettres ouïghoures.....	75
4.7	Création d'une méthode d'entrée virtuelle au niveau des navigateurs.....	77
4.8	Conversion multi-écriture	78
4.9	Evaluation et développement d'un dictionnaire en ligne.....	82
4.9.1	Affichage des lettres ouïghoures	82
4.9.2	Interface	82
4.9.3	Saisie des mots à rechercher	83
4.9.4	Recherche et présentation du résultat	84
4.9.5	Correction d'orthographe	85
4.9.6	Outil de recherche pour les webmasters.....	88

CHAPITRE 5. SUFFIXES VERBAUX DU OUÏGHOURE..... 92

5.1	Introduction	92
5.2	Consonnes	95
5.3	Voyelles.....	96
5.4	L'harmonie des consonnes.....	97
5.5	Harmonie vocalique.....	99
5.6	Règles euphoniques dans la suffixation	101
5.7	Morphologie	102
5.8	Règles de suffixation et cas spéciaux	104
5.9	Les expériences et les résultats	105
5.9.1	Dictionnaire des racines	106
5.9.2	Banque de règles	107
5.9.3	Générateur lexical.....	108
5.9.4	Explorateur de suffixes	109

CHAPITRE 6. MORPHOLOGIE DE L'OUÏGHOURE..... 111

6.1 Compréhension et connaissance	111
6.2 Structure phonétique	113
6.2.1 Les voyelles	113
6.2.2 L'affaiblissement des voyelles.....	113
6.2.3 L'harmonie vocalique.....	114
6.2.4 Les consonnes.....	116
6.3 Structure syllabique.....	117
6.3.1 Composition syllabique	117
6.3.2 Règlement de segmentation syllabique.....	119
6.4 Structure morphologique	120
6.4.1 L'aspect agglutinant des langues.....	120
6.4.2 Type des éléments agglutinants	122
6.4.3 Composition des mots ouïghours	125
6.5 Constitution des noms.....	127
6.5.1 Suffixes dérivatifs de noms.....	127
6.5.2 Suffixes syntactiques de noms	131
6.5.2.1 Le nombre.....	131
6.5.2.2 La possession.....	132
6.5.2.3 Les cas	133
6.5.2.4 L'ordre des suffixes, de nombre, de possession et des cas	134
6.6 Radical.....	134
 CHAPITRE 7. EXTRACTION DE L'INFORMATION OUÏGHOURE	 137
7.1 Les outils d'extraction.....	137
7.2 Deux types d'extraction.....	138
7.3 Affichage des caractères	139
7.4 Segmentation et traitement linguistique	140
7.5 Extraction des entités nommées.....	143
7.5.1 Un modèle pour traiter l'ouïghour	144
7.5.2 Problèmes d'harmonie vocalique et de l'affaiblissement des voyelles.....	146
7.5.3 Algorithme	147
7.6 Evaluation.....	148
 CHAPITRE 8. PARSEUR ET CORRECTEUR D'ORTHOGRAPHE MORPHOLOGIQUE.....	 150
8.1 Bref état des recherches	150
8.2 Qu'est ce qu'un parseur ?	152
8.3 Prétraitement.....	153
8.3.1 La segmentation en mots	153
8.3.2 Reconnaissance des écritures : ULY, UEY, USY	155
8.4 Identification de la racine	158
8.4.1 Recherche progressive	159

8.4.2 Recherche dégressive	160
8.4.3 Analyse de séquences modifiées	161
8.5 Vérification d'orthographe	165
CHAPITRE 9. CONCLUSION ET TRAVAUX FUTURES	168
CHAPITRE 10. BIBLIOGRAPHIE	173
CHAPITRE 11. ANNEXES	179
Annexe 1. Carte des groupes ethniques de l'Asie Centrale	180
Annexe 2. Alphabet turco - runique (VIème - IXème)	181
Annexe 3. Alphabet ouïghour ancien (Xème-XVIIIème)	183
Annexe 4. Table des alphabets de langues turciques.....	184
Annexe 5. Abréviations	185
Annexe 6. Suffixes ouïghours	187
Annexe 7. Lettres ouïghoures et leurs valeurs Unicode conventionnées.....	207
Annexe 8. Groupes des suffixes verbaux	209
Annexe 9. Liste des verbes primitive	211
Annexe 10. Tableau des alphabets ASU, LSU, CSU	216
Annexe 11. Arbre des langues turciques	218

Chapitre 1. Introduction

Avant d'aller plus loin, nous citons Mohamed Hatem HADDAD « *Dès l'invention des ordinateurs les hommes sont à la recherche d'une manière efficace de gérer, de stocker, de diffuser et de rechercher l'information. Plusieurs méthodes et techniques de gestion et de traitement d'information ont été développées. Aujourd'hui, nous pouvons estimer que nous sommes à un haut niveau d'informatisation grâce au développement et à la maîtrise de la technologie (soit celle des matériels, soit celle de la communication, soit celle de la construction des logiciels ou soit celle de la gestion et du traitement de l'information) dont l'Internet est un exemple flagrant.* » (M. H. HADDAD, 2002)

Or, bien que nous assistions à cette évolution remarquable, dans les domaines de l'informatique, de la communication et de l'information, nous nous rendons compte qu'il faut encore beaucoup progresser dans le domaine du traitement de l'information ouïghoure. C'est là notre souci. Le domaine de l'information est très vaste. Les problèmes que nous allons poser et étudier dans cette recherche concernent traitement automatique de la langue ouïghoure. Analyse linguistique est une des méthodes la plus utilisés pour le traitement automatique des langues. Notre objectif est d'étudier cette méthode et de l'appliquer dans le développement d'un moteur de recherche et correcteur orthographique pour l'ouïghoure. Avant d'entrer aux détails, nous commençons par expliquer le mot clé «ouïghour».

1.1 Les révolutions de l'informatisation ouïghoure

1.1.1 Qu'est que l'ouïghour¹ ?



L'ouïghour est une langue appartenant au groupe des langues turques de la famille des langues altaïques. Les Ouïghours constituent aujourd'hui la plus importante (officiellement 8,677,400 ², 19,860,000 ³ non officiel) des minorités nationales reconnues de la «région autonome» que la

¹ Un extrait de ce texte avait été ajouté sur Wikipédia et diffusé sur internet par l'auteur, voir <http://fr.wikipedia.org/wiki/Ou%C3%AFghour> ou rechercher sur internet.

² Le 5ème recensement de population, 2 avril 2001.

³ Ismail Cengiz, « Doğu Türkistan'ın 949-984 arasındaki nüfus değişimi II », Doğu Türkistan'ın Sesi, c. IV, sa. 31, Istanbul 1991, s.13-14.

Chine populaire appelle le Xinjiang (Nouveau Territoire) depuis 1949, et que les Ouïghours appellent le Turkestan Oriental [UAA⁴]. Les Ouïghours, peuple de langue turque dont le nom signifierait « alliance, unité », habitent traditionnellement en Asie centrale, dans les oasis du Takla-makan, les bassins de Turfan et de la Djoungarie et dans une partie du Ferghana. L'empire Ouïghour de Mongolie et les royaumes qui lui ont succédé en Asie centrale [Abdushükür Muhemmetimin, 2002] ont connu une brillante civilisation, jusqu'à leur absorption dans l'empire Mongol au XIII^{ème} siècle. Au cours de cette histoire, les Ouïghours ont adopté le Chamanisme, le Manichéisme, le Bouddhisme et le Nestorianisme pour finalement se convertir à l'Islam sunnite à partir du moment où les conquérants arabes battirent les Chinois en 751, ouvrant la voie à l'islamisation de l'Asie centrale. Sous l'influence de ces religions, les Ouïghours ont utilisé successivement et parfois de manière concurrentielle un grand nombre de systèmes d'écriture (turco-runique, brahmi, tokharien, soghdien) avant de développer sur la base de l'un des alphabets soghdiens leur propre système graphique, appelé depuis écriture ouïghoure ancienne. L'arrivée de l'Islam et l'absorption des régions de peuplement ouïghour dans l'empire turco-mongol musulman des descendants de Gengis Khan (empire Tchaghataï) ont amené le remplacement progressif de cette écriture par un alphabet arabo-persan.

Aujourd'hui, les Ouïghours de la « région autonome » ouïghoure de Chine, après 20 ans de romanisation sur base d'un système inspiré du Pinyin chinois, utilisent de nouveau l'alphabet arabo-persan (dorénavant « écriture ouïghoure »), mais sous une forme modifiée. La dernière réforme de l'alphabet ouïghour a été réalisée en 1983. Aujourd'hui, l'Arabe modifié est l'écriture officielle de la région « autonome » Ouïghoure de Chine, qui se compose de 24 consonnes et 8 voyelles. Elle comporte 27 lettres arabes modifiées et 5 lettres persanes.

1.1.2 Informatisation et langue ouïghoure

Depuis la première utilisation de l'informatique dans une maison d'édition au Xinjiang en 1984, la phase d'étude de faisabilité de logiciels pour le traitement des langues des minorités, a commencé à avoir du succès. On peut diviser ces périodes de révolution technologique en trois époques distinctes:

⁴ <http://www.uyghuramerican.org>

- **Traitement de texte ouïghour et développement de logiciel de logiciel de publication sous DOS (1984-1994)**

La mission principale dans cette étape était d'analyser et de comprendre la structure informatique. Des logiciels de méthode d'entrée de caractères ouïghours et de systèmes de publication multiécriture pour le gouvernement et les maisons d'édition ont été développés durant cette époque. L'intérêt de ces logiciels est qu'ils permettent de saisir des caractères issus de différents alphabets qui seront reconnus par l'ordinateur. Les résultats obtenus représentent maintenant une base pour les projets de traitement des logiciels multiécritures.

- **Études sur la plate-forme Windows (1994-1999)**

Le Groupe National 863⁵ de l'Université du Xinjiang a commencé de faire des recherches sur Windows en 1994 et a réussi à développer une version localisée en ouïghour de Windows 3.2 en juillet 1996. Le point de vue technique de cette innovation était d'adapter la version arabe de Windows 3.1 en faisant des changements de fonctions de contrôleur d'entrée-sortie, de localiser la disposition des touches du clavier, et de remplacer des glyphes de polices arabes afin d'implémenter des traitements multi-écritures.

- **Développement des logiciels d'application et réseaux (1999 - aujourd'hui)**

L'avancement dans plusieurs disciplines a apporté des résultats favorables à la croissance des technologies de traitement de l'information ouïghour. Les innovations les plus appréciées de cette époque ont été : une multitude de méthodes d'entrée⁶ pour Win 98/2000/NT/XP/Vista, de logiciels d'enseignement assisté par ordinateur, de logiciels orientés par objet⁷, de sites internet, de polices⁸ Unicode, de convertisseurs⁹ d'écriture

⁵ Le Groupe National 863 est spécialisé dans le domaine de la recherche scientifique de la haute technologie. (sorte de « CNRS chinois »)

⁶ qui n'utilisent malheureusement pas un standard de codage de glyphes dans la police, à cause de la difficulté d'affichage correct des formes (initiale, médiale, finale et isolé) des 3 caractères ouïghours, et manque de support des caractères ouïghours sous Windows 95/98 etc. En plus, il était plus facile de remplacer des glyphes arabes (dans la police) avec des glyphes ouïghours et de contrôler la disposition de la méthode d'entrée existante que les inventer à partir de zéro. Comme il n'y avait pas d'accord commun pour le remplacement des glyphes arabes avec des glyphes ouïghours, chacun a développé sa propre méthode. Par conséquent, il y a eu un problème d'incompatibilité des nombreuses polices créées par différentes personnes. Ainsi, les méthodes d'entrée et les logiciels doivent porter leur propre police. Il me semble que l'autorité régionale pour la standardisation aurait dû jouer un rôle important pour résoudre ce problème.

⁷ Purpose-oriented software.

ouïghoure vers UKY¹⁰, de dictionnaires bilingues¹¹, de logiciels de correction orthographique¹², UOCR¹³, la seule méthode d'entrée¹⁴ pour Win 98/2000/NT/XP/2003 qui utilise Unicode et un éditeur de texte¹⁵ Unicode. Une méthode d'entrée pour le ouïghour est disponible dans Windows Vista¹⁶. Mais, le traitement automatique de la langue ouïghoure n'a pas encore commencé.

1.2 Bref état des recherches

Les résultats obtenus jusqu'à nos jours pour le traitement de l'ouïghour sont loin d'être satisfaisants, ou bien ils sont très limités dans un cadre d'application pour la publication. Il n'y a pas d'outil qui peut faire l'analyse linguistique de l'ouïghour, ni de méthodes théoriques disponibles pour l'étude de la morphologie et de la syntaxe. Pour obtenir des prédictions profitables pour la langue ouïghoure, le meilleur moyen est de trouver, d'adapter et d'appliquer les ressources des langues proches.

Parmi toutes les langues agglutinantes de la même famille altaïque que l'ouïghour il y a par exemple: l'azéri¹⁷, le kazakh¹⁸, le kirghiz¹⁹, l'ouzbek²⁰, le tatar²¹, le turc et le

⁸ La première police ouïghoure qui utilise Unicode a été créée par Waris Abdukerim Janbaz en Juin 2002. Voir la dissertation pour obtenir le grade de M.Sc. chinois (équivalent à Bac+8 français), publiée en 2002 en chinois, disponible à la demande.

⁹ Par Waris Abdukerim Janbaz, téléchargeable à l'adresse suivante : <http://www.oyghan.com>.

¹⁰ «*Uyghur Kompyutér Yéziqi*» ou écriture informatique ouïghoure qui utilise l'alphabet latin. Nous consacrons un chapitre pour expliquer qu'est-ce UKY.

¹¹ Dictionnaire bidirectionnel chinois - ouïghour, voir: www.uyghursoft.com.

¹² Par le Groupe National 863.

¹³ Logiciel de reconnaissance optique des caractères ouïghours, par l'Université du Xinjiang avec la coordination de l'Université de Qinghua en août 2004.

¹⁴ Par Waris Abdukerim Janbaz, avril 2003, téléchargeable à l'adresse suivante : <http://www.oyghan.com>.

¹⁵ Par Gheyret Toxti Kenji et Muhammad Abla, téléchargeable à l'adresses suivantes : <http://uyghuredit.oyghan.com> et <http://www.bilik.cn>.

¹⁶ Un bug concernant le code de la lettre “ف” a été découvert après l'annonce de Windows Vista. Correction est prévu pour la future version.

¹⁷ Langue appartenant au groupe sud-ouest, ou Oghuz, des langues turques et très proche du turc parlé en Turquie. Elle est parlée par environ 10 millions de locuteurs répartis de part et d'autre de la frontière entre l'ancienne URSS et l'Iran

¹⁸ Langue turque parlée par les Kazakhs, peuple turco-tatar habitant la République du Kazakhstan et également en Région autonome ouïghour du Xinjiang au nord-ouest de la Chine.

turkmène²². La langue la plus proche de l'ouïghour est l'ouzbek, et celle qui est la plus avancée dans le domaine du TALN²³ est le turc. Certains modèles proposés par l'Université de Sabanci²⁴ et l'Université Bilkent²⁵ ont donné de bons résultats pour le turc. Grâce à la similarité de la grammaire et la ressemblance de la structure morphologique de l'ouïghour et du turc, nous pouvons bénéficier de l'expérience de nos prédécesseurs. Les méthodes utilisées pour la segmentation, l'annotation et l'extraction d'information du turc permettent cependant de se rapprocher au traitement de la morphologie de l'ouïghour.

1.3 Problématique de la recherche

L'arrivée du « World Wide Web » dans les oasis du Takla-makan et la grande utilisation d'Internet, ont été accompagnées d'une augmentation en corollaire de la demande de nombreux systèmes d'informatisations. L'analyse de la morphologie de l'ouïghour est une source de renseignements indispensable au développement de nouveaux logiciels qui sont enregistrés au sein de tous les systèmes du TALN. Depuis des années 2000, la demande d'étude sur les domaines suivants se multiplie :

- Développement des polices basées sur l'Unicode

¹⁹ Langue turque du groupe nord-ouest parlée par les Kirghiz, groupe nomade vivant en Kirghizistan, en Ouzbékistan et dans les montagnes du Badakhshan en République du Tadjikistan, Communauté des États indépendants, également en Région autonome ouïghour du Xinjiang au nord-ouest de la Chine.

²⁰ Langue turque parlée par les Ouzbek, peuple vivant en République d'Ouzbékistan, et en Région autonome ouïghour du Xinjiang au nord-ouest de la Chine. Cette langue est classée avec l'ouïghour dans le groupe sud-est des langues turques, descendant direct du Tchaghataï.

²¹ Langue turque parlée en république autonome des Tatars (Tatarija, Tatarie) Fédération de Russie, et également en Roumanie, Bulgarie, Turquie et en Région autonome ouïghour du Xinjiang au nord-ouest de la Chine. Les dialectes tatars sont nombreux et comprennent notamment : le tatar de Crimée, le tatar de la Volga, le tatar de Kazan et le tatar de Lituanie.

²² Langue turque parlée principalement au Turkménistan, ainsi que dans l'ouest de l'Ouzbékistan, le sud du Kazakhstan, le nord de l'Iran et de l'Afghanistan. Le turkmène appartient au groupe Oghuz des langues turques et est très proche de l'azéri et du turc de Turquie.

²³ Traitement automatisé du langage naturel.

²⁴ Human Language and Speech Technologies Laboratory, Faculty of Engineering and Natural Sciences (www.sabanciuniv.edu).

²⁵ Department of Computer Engineering and Information Science, Ankara, 06533 Turkey (www.bilkent.edu.tr).

- Méthodes d'entrées
- Logiciel de correction orthographique de l'ouïghour
- Moteur de recherche ouïghour
- Logiciel de traduction automatique.
- Logiciel de traduction assistée par ordinateur.
- Système d'extraction d'information monolingue/multilingue
- Banque de données terminologiques monolingue/multilingue
- Analyseur d'archives documentaires
- Etc.

Une démarche d'étude linguistique est inévitable dans toutes les procédures de réalisation des logiciels ci-dessus. Notre étude s'intéresse à la conception d'analyse morphologique de l'ouïghour, quelles que soient les applications auxquelles elles sont destinées. Plus particulièrement, comment créer des polices ouïghoures ? Comment entrer et afficher des lettres ouïghoures ? comment faire l'extraction d'information ? Quelles sont les problématiques morphologiques à concerner ? Quelles sont les particularités de la langue ouïghoure ? Comment résoudre les problèmes techniques rencontrés ?

«Il existe bien des difficultés à surmonter pour créer un système de RD efficace : rapidité de l'indexation et de la recherche, taille de l'index, robustesse, fiabilité, efficacité, etc. Mais les problèmes les plus difficiles ne correspondent pas à de la technique pure. Ils sont liés aux propriétés même des langues.» (Claude de Loupy, 2001). Nous évoquerons, du point de vue de la linguistique et de l'informatique, certaines des difficultés les plus importantes liées au traitement automatique de la langue ouïghoure dans une situation où les systèmes d'exploitation ne supporte pas la langue ouïghoure.

1.4 Plan de thèse

Dans cette étude, nous décrirons comment concevoir des algorithmes de traitement automatique de langue adaptés pour d'autres langues que les langues courantes, et plus précisément pour l'ouïghour.

Tout d'abord, nous faisons une comparaison de vocabulaire entre 8 langues turques, en présentant la particularité de l'ouïghour. Ensuite, nous présenterons les caractéristiques

et la structure morphologique de la langue ouïghour en faisant une étude des suffixes de dérivation. Enfin, nous décrirons les étapes afin d'implémenter des outils informatiques en tenant compte notamment des problèmes techniques et linguistiques posés au traitement de l'ouïghour.

Notre approche montrera qu'à partir d'un outil et de telles techniques, il est possible d'utiliser une approche purement linguistique pour faire de recherche de l'information, d'analyse morphologique et de correction orthographique. La partie évaluation de notre recherche montre quelle source de données a été choisie pour extraire les données sur les interactions. Nous proposerons des algorithmes qui seront utiles pour la syllabisation, le parseur et pour la suggestion orthographique. Notre étude se développera dans deux grandes étapes :

- **Problèmes traités au niveau de préparation**

- Création d'une méthode d'entrée pour taper des caractères en alphabet arabe
- Etude sur les problématiques de multi-écriture
- Développement des polices pour le codage et l'affichage des caractères arabe
- Implémentation d'un convertisseur multidirectionnel
- Préparation d'un dictionnaire

- **Étude linguistique**

- Etude sur la nature agglutinante: une langue agglutinante dans laquelle les mots sont longs et complexes car ils se composent d'une série de morphèmes concaténés, comme un chapelet. E.g. *Küchlendürelmeywatqanliringlarningkidinmu?* (racine: *küch*, + 14 suffixes)
- Un générateur lexical basé sur les règles morphologiques de suffixation de l'ouïghour
- Un analyseur morphologique qui permet d'expliquer les caractéristiques d'agglutination des mots suffixés

Chapitre 1. La gestion du savoir

L'objectif de ce chapitre est d'abord de définir les concepts concernés pendant l'extraction de l'information, ensuite de décrire les caractéristiques de la langue afin de comprendre les problématiques, et enfin de comparer la morphologie des langues turques.

1.1 Définition et enjeux

1.1.1 La recherche

Nous annexons ici la définition de Christine MICHEL : « *Les questions usuelles que les expérimentateurs se posent pour évaluer la qualité de la recherche d'information sont :*

- *Quelle est la quantité d'information minimale que l'utilisateur est en droit de demander suite à une question ?*
- *Quelle est la « quantité d'information » contenue dans l'ensemble des réponses proposées, ou dans un texte particulier, en fonction du problème soulevé ?*
- *Comment mesurer l'information manquante ?*

Les tests usuels d'évaluation des bases de données sont basés sur des collections tests et mesurent le rappel et la précision de la réponse en référence à un ensemble de réponses « pertinentes ». La notion de document pertinent peut varier. Il peut être défini comme un document proche de la question ou bien utile pour l'utilisateur. Tout comme le concept de qualité informationnelle, le concept de pertinence varie suivant l'étude menée, le problème soulevé et l'utilisateur lui-même. »²⁶

1.1.2 L'information

L'information est une ressource stratégique — tout le monde le répète et on ne manque pas de données, de nos jours, qui le confirment. D'un point de vue scientifique, l'information apparaît comme un sujet vague et incohérent. « *Le mot information a des définitions multiples et ambiguës. La définition du Larousse est tout à fait significative. Elle se décompose en plusieurs sous définitions selon les critères suivants :*

²⁶ Christine Michel « Evaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogènes » <http://www.recodoc.univ-lyon1.fr/theseCMichel.pdf>

- *Le critère Action : l'information est l'action d'informer, de se mettre au courant d'événements.*
- *Le critère Etat : l'information est une nouvelle, un renseignement que l'on communique ou que l'on obtient.*
- *Le critère Connaissance : l'information est un ensemble de connaissances acquises sur quelqu'un ou sur quelque chose.*
- *Le critère Contenu : l'information est le contenu proprement dit des messages transmis.*
- *Le critère Contenant : l'information est un signal par lequel un système donne connaissance de sa position à un autre » [Mohamed Hatem HADDAD, 2002].*

Selon un autre critère : « *L'information est émission, réception, création, retransmission de signaux groupés oraux ou écrits, sonores, visuels ou audiovisuels en vue de la diffusion et de la communication d'idées, de faits, de connaissances, d'analyses, de concepts, de plans, d'objets, de projets, d'effets de toute sorte dans tous les domaines, par un individu, par des groupes d'individus ou par un ou plusieurs organismes agissant ou rétroagissant ainsi sur leur environnement immédiat, proche ou lointain, et dont le but est de déclencher éventuellement des processus dialectiques plus ou moins amples alimentant l'échange, base naturelle et indispensable de l'animation de la vie sociale.* »²⁷

Lorsqu'on a trop d'information, il est difficile de repérer les éléments importants parmi la masse de documents existants. Par conséquent, on a besoin d'un système ou d'outils d'extraction afin de trouver le point culminant de l'information qui nous intéresse. De ce fait, dans cette étude, nous donnons la définition suivante : **l'information** est la partie profitable, dont les utilisateurs ont besoin en priorité et en particulier, dans un document archivé.

²⁷ <http://www.olats.org/schoffer/definfo.htm>

1.1.3 Définition d'extraction d'information

Depuis que les ordinateurs sont apparus, des milliards d'informations y ont été enregistrées dans plusieurs bases de données, dans divers domaines de connaissances et sous diverses formes (textes, images, son, vidéo etc.). Etant donné que les ressources informationnelles sont de plus en plus accessibles aux utilisateurs personnels, le principal problème aujourd'hui est de savoir comment accéder à l'information dont on a besoin.

« L'extraction d'information consiste à remplir automatiquement des formulaires ou une banque de données à partir de textes écrits en langue naturelle. Elle s'oppose classiquement à la recherche documentaire qui vise à retrouver dans une base de documents un ensemble de documents pertinents au regard d'une question. L'extraction met en oeuvre une analyse du texte pour interpréter et construire une représentation formelle qui permettra d'apporter automatiquement des réponses précises à l'utilisateur » [Ph. Bessière & A. Nazarenko & C. Nédellec, 2002²⁸]. Lorsque l'analyse est automatique, la machine extrait l'information pertinente contenue dans un objet en s'appuyant sur les éléments textuels de l'objet. La tâche est difficile car la cible « langage naturel » est complexe et ambiguë. A un problème, déjà un énoncé textuel, s'ajoute le problème de la reconnaissance linguistique automatique du texte. Les linguistes ont été nombreux à travailler sur la résolution des problèmes de traitement plus ou moins automatique du texte dans le cadre du traitement linguistique informatisé. A travers les problèmes linguistiques, les linguistes veulent comprendre comment se fait la connaissance du sens.

L'extraction d'information consiste donc à identifier de l'information bien précise d'un texte en langue naturelle mais aussi à pouvoir la représenter sous forme structurée. L'extraction d'information peut être une partie d'un système de recherche d'information (dorénavant : SRI) chargée de recevoir une requête, de la traiter, de déterminer la similitude entre la demande d'information contenue dans la requête et les items d'information de la base de données.

²⁸ <http://www-leibniz.imag.fr/SICLAD/Caderige/Articles/Cide-caderige.pdf>

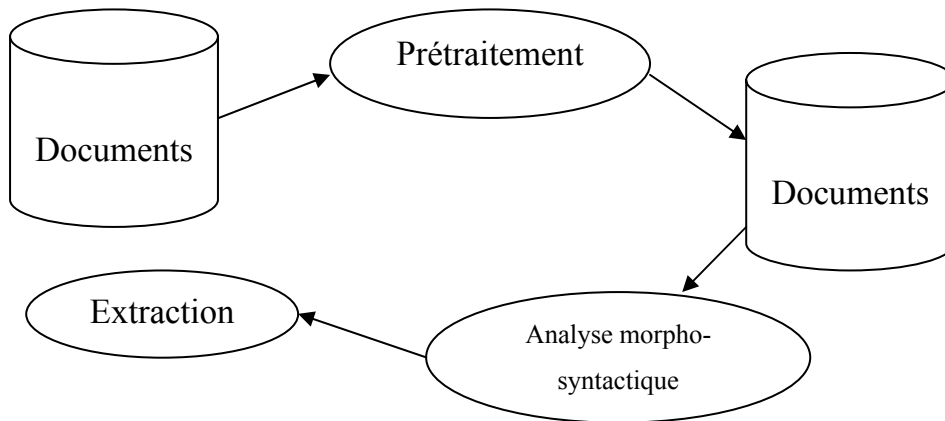


Figure 1. Processus d'extraction d'information

L'extraction d'information est donc une tâche qui consiste à extraire de l'information structurée à partir d'un document textuel.

1.1.4 Définition d'un Système de Recherche d'Information

Donnons un exemple d'abord : dans un ensemble de données nous cherchons :

- « Victor Hugo est né »

Un article qui parle de la date de naissance de Victor Hugo (repérage de texte)

- « Carte de la France »

Une carte de la France (repérage d'image)

- « Elle est à toi cette chanson »

Une chanson de Georges Brassens (repérage de son / music)

- « Les Guignols de l'Info »

Un émission de Canal + (repérage de vidéo)

Il s'agit d'extraire de l'information spécifique (texte, image, son, vidéo) et pertinente d'un ensemble de données collectées afin de fournir une information élaborée et synthétique. Reste à déterminer comment et par quel moyen on pourra récupérer d'un texte une phrase qui traite d'un sujet donné.

« Un SRI est un système composé d'une part d'un module chargé du traitement, de l'indexation et du stockage de l'information. Ce module construit, à partir du traitement de l'information, une structure de données organisées de manière à permettre l'accès rapide à l'information. D'autre part, il est composé par un module, appelé interface, qui sert à

interagir avec les usagers, dotée des mécanismes de sélection d'information orientés par les requêtes formulées par les usagers. Nous distinguerons, donc, ce qu'est un SRI et ce qu'est un système d'extraction d'information.» [Mohamed Hatem HADDAD, 2002].

1.1.5 Rappel et Précision

« Un système de recherche d'informations est évalué en observant ses réponses par rapport à ce que l'utilisateur souhaite retrouver. Les mesures généralement utilisées à cet effet sont les mesures de rappel et de précision, comprises dans l'intervalle [0,1]. » [Support de cours du CRIM / INALCO 2003-2004, «l'extraction d'information », par Thierry Poibeau, Page 10].

1.1.6 Indexation

Selon l'approche documentaire, AFNOR [NF Z47-102, p.512, 1996] *« l'indexation est l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document, c'est-à-dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse.»*. L'indexation constitue selon [WALTER, 1999] l'une des étapes de l'analyse documentaire au même titre que la lecture ou le résumé.

L'indexation automatique est basée sur l'indexation contrôlée de l'approche documentaire. Ainsi, *« l'indexation est un processus destiné à représenter par les éléments d'un langage documentaire ou naturel des données résultant de l'analyse du contenu d'un document ou d'une question. Elle représente le contenu intellectuel d'un document par des mots-clés (mots ou expressions) jugés significatifs de ce contenu par analyseur. Leur collecte, manuelle ou automatisée, donne lieu à l'établissement d'une liste d'indices de classification qui permet une recherche efficace des informations contenues dans un fonds de document et d'indiquer rapidement, sous forme concise, le contenu d'un document »*²⁹.

Les procédures d'extraction automatique sont des préalables de l'indexation automatique. Deux approches d'extraction automatique — linguistique et statistique, nous permettent de faire des traitements automatiques aux niveaux morphologique et sémantique.

²⁹ Glossaire du département documentation de l'UQAM,
http://www.bibliotheques.uqam.ca/bibliotheques/serv_techniques/analyse/politiques/POL_glossaire.HTML

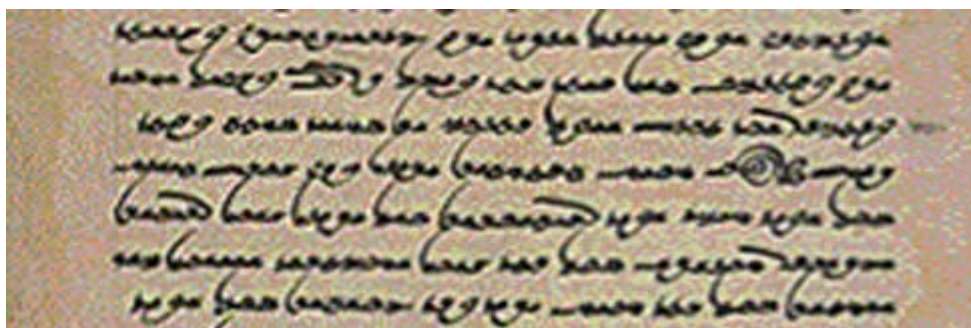
Dans notre étude, nous présenterons l'approche linguistique d'extraction automatique du niveau de combinaison des mots, c'est-à-dire au niveau morphologique.

1.2 Particularités de l'ouïghour

1.2.1 Écritures ouïghoures

L'écriture ouïghoure est très ancienne. Très tôt les ouïghours semblent avoir disposés de structures administratives complexes [Qurban Weli, 1986] : il est donc naturel qu'on y ait développé toute une série de moyens servant à conserver la mémoire des faits et des idées et, de ce point de vue, l'écriture ne représente qu'une des techniques expérimentées dans le monde ouïghour. Les Ouïghours ont utilisé différentes écritures à différentes époques. Les plus importantes sont :

- **l'écriture sogdienne (V^{ème} – X^{ème})**, dérive de l'écriture syriaque au début du V^{ème} siècle. Ses plus anciennes traces ont été découvertes dans les régions de Khotan, Turfan, Kucha, Qumul, Qarasheher et Dunhuang. On peut aussi trouver cette écriture dans les inscriptions de la région de l'Orkhon qui appartiennent à l'époque de l'empire ouïghour (745-840).



Exemple d'écriture : extrait³⁰ de «Le Sutra des causes et des effets du bien et du mal »

Dunhuang, VIII^e – IX^e siècle, Manuscrit sogdien, encre sur papier. H : 700 cm ; L : 26 cm. Mission Pelliot, 1906-1909, Ming Öy, Paris, Bibliothèque Nationale de France

³⁰ Oasis perdues des Routes de la Soie, Serge Viallet. © UNESCO, Beta System, R.M.N. Paris 1995. (Format CDROM).



Extrait³⁴ de « Lettre officielle en ouïghour »

Khotan, ca 948

Manuscrit, encre sur papier, H : 28 cm ; L : 37 cm

Mission Pelliot 1906-1908, Ming Öy

Paris, Bibliothèque Nationale de France

- **l'écriture arabo-persane**³⁵ (XI^{ème} -**aujourd'hui**). Après l'introduction de l'Islam, les Ouïghours ont adopté l'écriture arabe qui a cohabité avec l'écriture ouïghoure ancienne jusqu'au XVIII^{ème} siècle. Le « Grand dictionnaire des langues turques » (Türkiy Tillar Diwani) de Mahmud Kashgari a été écrit au XI^{ème} siècle en écriture arabe. S'adaptant progressivement à la réalité phonétique des langues turques, cet alphabet a évolué pour donner naissance à un système d'écriture arabo-persane capable de transcrire tous les éléments étymologiquement fort différent de la grande langue littéraire commune à tous les peuple d'Asie centrale jusqu'au début du XX^{ème} siècle appelée tchaghataï.

Depuis le début du XX^{ème} siècle, une série de réformes successives visant à rendre l'écriture plus proche de la phonétique de l'ouïghour et a lui faire refléter plus fidèlement les sons de la langue parlée a abouti à la mise au point progressive de l'écriture actuellement en usage parmi les Ouïghours hors du territoire de la Communauté des États Indépendant (ex Union soviétique, où les Ouïghours utilisent toujours un alphabet dérivé du cyrillique).

- **L'alphabet cyrillique**³⁶ **ouïghour**, créé pour des raisons politiques au cours des années 40 pour les Ouïghours de l'ex Union Soviétique, qui avaient utilisé un système romanisé

³⁴ Oasis perdues des Routes de la Soie, Serge Viallet. © UNESCO, Beta System, R.M.N. Paris 1995. (Format CDROM).

³⁵ Voir annexe IV, écritures ouïghoure

³⁶ Voir annexe IV, écritures ouïghoure

pendant près d'une vingtaine d'années auparavant, il comporte aussi des lettres uniques pour les sons propres au ouïghour.

- **le Pinyin**³⁷ (**Yéngi Yéziq ou nouvelle écriture**), imposé pour des raisons politiques par les autorités de Pékin pour remplacer l'alphabet arabo-persan (alors rebaptisé Kona Yéziq ou ancienne écriture), il a dû céder de nouveau la place à celui-ci en 1982. Il utilise l'alphabet latin, avec les valeurs phonétiques qu'il représente pour la transcription du chinois, et un certain nombre de lettres uniques créées spécialement pour transcrire des sons ouïghours. Ces nouvelles lettres représentent un obstacle à l'informatisation de l'ouïghour.

- **le UKY**³⁸ (**« Uyghur Kompyutér Yéziqi » ou écriture latine ouïghour**), *« créé par un comité d'utilisateurs, d'informaticiens et de linguistes en 2001 comme forme unique de romanisation dans le contexte informatique, il utilise l'alphabet latin avec la valeur des lettres la plus commune dans la communauté internationale et en limitant au minimum le recours à des signes diacritiques. Il vise ainsi à résoudre les problèmes d'entrée de l'ouïghour écrit dans les alphabets actuellement en usage (arabo-persan, pinyin, cyrillique ou même turc) sans pour autant ajouter la difficulté d'apprentissage d'un système de transcription trop complexe. Il faut préciser ici que jusqu'à l'apparition du UKY, les Ouïghours se trouvaient contraints d'avoir recours au chinois ou de créer leurs propres transcriptions, souvent peu systématiques, pour utiliser l'ordinateur ou communiquer sur le web. Ainsi, la lettre q pouvait souvent transcrire le son [tch] aussi bien que le son [q] et la lettre x représentait le plus souvent le son [ch].*

Les autorités de la « Région autonome » ouïghoure ont décidé de permettre l'utilisation du UKY à titre expérimental, ce qui lui a permis de prendre un essor rapide. Aujourd'hui, certains sites Internet ouïghours et quelques livres ouïghours en ligne utilisent cette transcription. Il existe plusieurs logiciels³⁹ pour convertir entre l'UKY, le cyrillique ouïghour et l'écriture arabo-persane.

Soulignons que l'introduction de l'UKY ne vise pas à mettre fin à l'existence de l'écriture arabo-persane. En effet, il existe de plus en plus de programmes informatiques en

³⁷ Voir annexe IV, écritures ouïghoure

³⁸ Voir annexe IV, écritures ouïghoure. Un extrait de ce texte avait été ajouté sur Wikipédia et diffusé sur internet par l'auteur, voir <http://fr.wikipedia.org/wiki/Ou%C3%AFghour>.

³⁹ Voir site personnel de l'auteur: <http://www.oyghan.com>

écriture arabo-persane mais les réalités socio-économiques des régions ouïghoures se combinent aux difficultés techniques pour en limiter la portée ».

1.2.2 Caractéristiques techniques de l'écriture

Aujourd'hui, seule l'écriture arabo-persane modifiée [voir annexe IV, et ci-dessus] est l'écriture officielle de la «région autonome» ouïghoure de Chine. Elle se compose de 24 consonnes et 8 voyelles qui permettent de transcrire tous les sons de la langue ouïghoure. Pour ce faire, il a été nécessaire de créer un certain nombre de lettres qui n'existent pas dans l'alphabet arabe ou persan, ni dans aucun de leurs alphabets dérivés, ce qui occasionne des problèmes au niveau de l'informatisation de l'ouïghour.

1.2.2.1 Nombre de signes

L'écriture ouïghoure se compose de 24 consonnes et 8 voyelles. Chaque une des lettres a de 2 à 8 glyphes différents (initiale, médiane, finale, isolée). Certaines formes des voyelles ouïghoures se forment de deux glyphes. Après la dernière réforme en 1983, l'alphabet ouïghour est constitué de 106 glyphes, 20 ligatures⁴⁰.

Un glyphe est une image utilisée pour représenter les différentes formes des lettres. Dans une police, chacune des formes des lettres est appelée des glyphes. Une ligature est une composition de deux ou plusieurs glyphes des lettres. Dans le cadre de l'écriture arabe, deux ou plusieurs glyphes des lettres s'associent pour former une ligature. Dans plupart des cas, les ligatures sont disponibles seulement dans l'extension du jeu de caractères des polices. Par exemple : Arabic Presentation Forms-A (FCB0 - FDFF41). En arabe, plusieurs glyphes de plusieurs lettres forment une ligature. Par exemple :

الله = la forme isolés de la lettre *alef* «ا» + la forme initiale de la lettre *lam* «ل» + la forme médiane de la lettre *lam* «ل» + la forme finale de la lettre *heh* «ه».

Une des plus courantes ligatures ouïghoure est « ۋ ». Quand la lettre « ۋ » (0627) s'interpose avec la lettre « ۋ » (0626), ils deviennent « ۋ » (FBEA) en associant ses deux

⁴⁰ voir http://www.oyghan.com/unicode/Unicode_FAQ_fichiers/image005.gif

⁴¹ valeur Unicode en hexadécimale.

glyphs: « ۞ » *ARABIC LETTER YEH WITH HAMZA ABOVE INITIAL FORM*⁴² (FB8B) et « ۞ » *ARABIC LETTER ALEF FINAL FORM*⁴³ (FE8E).

Les voyelles : ئا، ئه، ئې، ئى، ئو، ئۆ، ئۇ، ئۈ

Les consonnes: ب، پ، ت، ج، چ، خ، د، ر، ز، ژ، س، ش، غ، ف، ق، ك، گ، گ، ل، م، ن، ه، و، ي

1.2.2.2 Type d'écriture

Les écritures ouïghoures citées plus haut possèdent au moins une caractéristique essentielle commune : ce sont des écritures alphabétiques phonétiques.

1.2.2.3 Direction de l'écriture

L'écriture ouïghoure, comme la plupart des écritures sémitiques⁴⁴, s'écrit de droite à gauche.

1.2.2.4 Caractères de ponctuations et des chiffres

L'ouïghour utilise les chiffres arabes, 0123456789, comme le français et plupart des langues du monde. Mais le point d'interrogation, virgule, point virgule sont différents ; à savoir : ؟ ، ، ؛ .

1.2.2.5 La spécificité de l'écriture ouïghoure

- Il n'y a pas de majuscule.
- L'écriture ouïghoure est cursive. Les voyelles peuvent s'attacher à la lettre d'avant ou rester sans se joindre. Toutes les consonnes et seulement deux voyelles (ئې، ئى) se lient à la lettre suivante, sauf quatre consonnes, à savoir : ئ، ژ، د، ر. Cette ligature a donné lieu à la création d'une belle calligraphie.
- Les lettres ouïghoures peuvent théoriquement avoir quatre formes différentes : à l'initiale, au milieu, à la fin du mot et en position isolée.
- Contrairement de l'arabe, il est toujours nécessaire d'écrire les voyelles.

⁴² Forme initiale de la lettre arabe « Yeh » avec « Hamza » dessus, voir www.unicode.org .

⁴³ Forme finale de la lettre arabe « alef », voir www.unicode.org .

⁴⁴ Famille chamito-sémitique (afro-asiatique) comprenant l'akkadien, l'arabe, l'araméen, l'éthiopien, le hébreu et le phénicien et parlée au nord de l'Afrique et en Asie moyen-orientale.

1.2.2.6 Identification de la langue ouïghour

La deux-lettres-code (2 letter code) attribuée par l'ISO pour l'ouïghour est « ug », et la trois-lettres-code (3 letter code) est « uig ». Grâce à ces identifiants, la langue ouïghour peut se distinguer des autres langues pendant le traitement textuel en informatique. Dans les balises HTML/XML, le paramètre de l'attribut *lang* permet d'indiquer la langue, l'attribut *dir* peut aligner la direction d'affichage des caractères. Par exemple :

```
<p dir="rtl" lang="ug" >ئۇيغۇر تىلى</p>
```

dir="rtl" indique la direction d'affichage, de droite à gauche; *lang="ug"* indique la langue du contenu du paragraphe, ouïghour.

1.2.2.7 Méthodes d'entrée de l'ouïghour

Il faut noter qu'aucun système d'exploitation ne contient de méthode d'entrée pour l'ouïghour. De ce fait, nombreux chercheurs indépendants ont réussi à réaliser des méthodes d'entrées différentes.

L'écriture ouïghoure utilise les méthodes suivantes sur Windows :

- Entrée directe par syllabes
- Entrée directe alphabétique
- Méthode *Copier-Coller*

Pour ceux qui ne connaissent pas l'alphabet ouïghour, on peut proposer :

- Méthode de conversion latin-ouïghour

Il est possible d'écrire l'ouïghour avec les lettres latines (voir UKY). Ainsi, par exemple, on peut écrire « Uyghur » qui est l'équivalent de « ئۇيغۇر ».

Comme on l'a mentionné dans le premier chapitre (voir paragraphe 1.1.2), malgré la multiplication des méthodes d'entrées, la compatibilité reste toujours un problème à résoudre jusqu'à la création de la première méthode d'entrée⁴⁵ Unicode, pour Windows, en 2004. Récemment, Microsoft a annoncé le projet d'ajouter une méthode d'entrée ouïghour à la nouvelle version du Windows Vista / 7.

⁴⁵ Par l'auteur, Waris Abdukerim Janbaz, disponible gratuitement à l'adresse suivant : <http://www.oyghan.com>

1.2.2.8 Stockage et bases de données

Le stockage de l'information multilingue et des caractères ouïghours pose de nombreux problèmes. Pour les nécessités du traitement (recherche de chaîne de caractères, tri, indexation, etc.) le texte codé en mémoire doit être constitué d'une suite de « lettres abstraites » stockées suivant des adresses croissantes dans l'ordre de lecture (c'est le cas pour tous les textes). Nous nous sommes donc trouvés dans l'obligation, pour ce document, de transgresser provisoirement les normes en stockant les fragments de texte dans l'ordre visuel et en codant les formes de présentation (formes liées). La base de données doit être capable d'épargner correctement les caractères Unicode pour implémenter l'algorithme bidirectionnel (appliqué aux écritures s'écrites de droite à gauche, ex : arabe, hébreu) de l'Unicode sur les navigateurs. Heureusement, la plupart des logiciels (version 2000 ou supérieure) de base de données supportent l'Unicode.

1.2.3 Extension territoriale et langues notées

Comme l'écriture ouïghoure est tirée de l'écriture arabe, elle devrait appartenir au grand domaine de l'écriture arabe. La prédominance de la langue écrite sur les langues parlées d'Asie Centrale a en effet rendu cela possible jusqu'au début du XX^{ème} siècle. Mais la création de langues ethniques turques (ouïghour, ouzbèk, kazakh, kirghiz, Karakalpak, tatar etc.) basées sur la langue parlée à partir du milieu des années 20 du XX^{ème} siècle, suivi d'un mouvement incessant de réformes d'écriture, avec la création de lettres propres à chacune de ces langues⁴⁶, a mis fin à cet état de fait.

Actuellement, l'écriture ouïghoure est utilisée en région autonome ouïghoure du Xinjiang (ou Turkestan Chinois) par les Ouïghours, ainsi que par les Ouzbèks et les Tadjiks du Xinjiang.

⁴⁶ Voir annexe 1 groupe d'ethnie majeur de l'Asie Centrale.

1.3 La spécificité de l'ouïghour

L'ouïghour est une langue appartenant au sous-groupe Huns⁴⁷ [Haji Nurhaji, *Qedimki Uyghurlar we Qaraxaniylar*, Page 121, 2001] du groupe Sud-est des langues turques de la famille des langues altaïques.

« En repartant de l'Europe, on trouve le groupe des langues finno-ougriennes, incluant d'une part finnois, estonien, live, tchérémisse, et, de l'autre, hongrois, ostiak. Si l'on y adjoint, d'un côté, le lapon, et, sur le Yénisséi, les langues samoyèdes, on obtient une famille plus vaste, celle des langues ouraliennes. Ces langues sont agglutinantes et offrent des traits tels que la position finale du verbe et l'harmonie vocalique, sorte d'assimilation à distance modifiant la voyelle du (ou des) suffixe(s) pour l'assimiler au moins partiellement à la voyelle radicale. Si celle-ci est d'avant (e, ö, ü) les voyelles figurant dans les différents affixes seront aussi nécessairement des voyelles frontales; et de même pour les voyelles d'arrière (a, o, u) »⁴⁸. Par exemple :

hongrois : *hazban* (dans la maison), est formé à l'aide de la racine *haz* et du suffixe locatif – *ban*.

turc : *evde* (dans la maison), est formé à l'aide de la racine *ev* et du suffixe locatif – *de*.

turc : *öyde* (dans la maison), est formé à l'aide de la racine *öy* et du suffixe locatif – *de*.

L'ouïghour partage les caractéristiques des langues ouralo-altaïques qui se distinguent des langues indo-européennes. A savoir :

- harmonie vocalique
- absence de genre (pas de masculin/féminin)
- agglutination (voir 3.5.1)
- adjectif épithète toujours avant le nom
- sujet toujours avant le prédicat
- verbe toujours en finale
- pas de verbe avoir
- chaque lettre correspond à un seul son

⁴⁷ Voir Huns, peuple d'origine de langue altaïque, venue en Europe aux IV^{ème} et V^{ème} siècle après J.C. sous la conduite d'Attila.

⁴⁸ Voir <http://www.ciep.fr/publications/genetique/genetique33.php>

Voir le site <http://www.ciep.fr/publications/genetique/genetique33.php>

L'ouïghour est une langue très ancienne qui a une structure complexe aussi bien sur le plan phonétique que morphologique ou syntaxique. Elle possède également un vocabulaire riche de 70,000 à 100,000 mots. L'ouïghour est composé de mots que des suffixes en grand nombre vont permettre de décliner et de conjuguer afin de construire la phrase et lui donner toute sa signification. Dans le 3^{ème} chapitre, nous présentons les caractéristiques de l'ouïghour en faisant une étude plus profonde de sa morphologie.

1.4 Un regard sur les langues turques

Les ouïghours ne représentent qu'un seul des peuples tures, mais il en existe bien d'autres. La classification de langues turques est très variée. Il y a au moins une dizaine de classifications, et la plupart d'entre elles font l'objet de conflits intérieurs et manquent de cohérence taxinomique. Le site d'Internet www.ethnologue.com donne une liste de 40 langues⁴⁹ turques. Par ailleurs, certains turcologues conçoivent que ce nombre peut être supérieur à 50. Parmi les classifications il y a des classifications selon la géographie, le nom du groupe de l'ethnie, la caractéristique lingo-génétique. Pour des réponses plus précises, consultez des ouvrages de G.Ramsted, V.V. Radlov, A.N. Samoilovich, N.A. Baskakov, V.A. Bogoroditskiy, M.T. Diachok, Mirsultan Osmanov.

Le site d'Internet www.ethnologue.com nous donne la classification de langues turques⁵⁰.

⁴⁹ Voir http://www.ethnologue.com/show_family.asp?name=Altaic.

⁵⁰ Attention : la source des documents informatique sur la Toile change très souvent d'adresse, et sont parfois insatisfaisants.

Il faut noter que les orthographes du nom de l'ethnie et de la langue sont très variées. Rien que pour l'orthographe de la langue ouïghour on peut trouver des variantes comme « ouïghour, ouïgour, ouïghour, uigur » en français, et comme « Uyghur, Uighur, Uygur, Uigur, Wighor, weiwuer » en anglais. Dans cette étude, nous utilisons la norme la plus fréquemment utilisée depuis des années 90, à savoir: « ouïghour » en français et « Uyghur » en anglais. Le nom du groupe de l'ethnie et de la langue est prononcé [yjɣyr] et est s'écrit « ئۇيغۇر » en ouïghour. L'utilisation de « gh » pour des noms propres qui ont le son [ɣ] (r comme roi) est courante en français. Par exemple : Afghan, Gharb, Maghreb.

L'ouïghour est une langue relativement facile à apprendre pour les turcophones. Il est difficile de donner un pourcentage exact des mots qui se ressemblent entre les langues turques sans faire une statistique de manière rigoureuse. Mais on peut dire que les ouïghours n'ont pas du tout de problème de compréhension de l'ouzbek, et ils arrivent à comprendre assez bien le sujet principale d'un discours en kazakh, kirghiz. En revanche, les ouïghours ont des fois besoin d'un interprète pour comprendre la langue de leur cousin géographiquement éloigné – le turc. Le vocabulaire ouïghour comporte 60 à 97 pourcent⁵¹ de mots co-existant dans les autres langues turques, ainsi que des emprunts arabes, persans, russes, chinois, anglais et d'autres langues.

Afin de mieux comprendre la morphologie des langues turques, nous ferons une petite analyse en donnant des exemples. Une étude parallèle de toutes les langues turques est une tâche qui prend beaucoup de temps. Ainsi, nous présenterons que 8 langues turques afin de faire une mini-comparaison des écritures⁵² et de la prononciation de certains mots choisis de manière aléatoire.

Azéri	Дин din	өлдүр- öldür-	мүһүм mühüm	көј göy	үч üç
Kazakh	дін (din)	өлтір- (öltir-)	маңызды (mangızdı)	көк (kök)	Үш (üşh)

⁵¹ Selon les estimations incertaines.

⁵² Voir l'annexe 4

Kirghiz	дин (din)	өлтүр- (öltür-)	маанилүү (maanilü)	көк (kök)	Үч (üch)
Tatar	Дин din	үтер- üter-	мөһим möhim	зәңҗар zənggär	өч öç
Turc	Din	öldür-	Önemli	gök	Üç
Turkmène	Дин din	өлдүр- öldür-	мөхүм möxüm	гөк gök	үч üç
Ouïghour	دىن (din)	ئۆلتۈر- (öltür)	مۇھىم muhim	كۆك (kök)	ئۈچ (üch)
Ouzbek	Дин din	ўлдир- o'ldir-	муһим muhim	кўк ko'k	уч uch
Français	Religion	tuer	important	bleu (et/ou le ciel)	Trois

Tableau 1. Une comparaison de noms, verbe, adjectif et de chiffre. Translittérations en écriture latine apparaissent après l'écriture officielle dans les parenthèses.

Azéri	-да, -дә -da, -də	-а, -ә, -ја, -јә -а, -ә, -ја, -јә
Kazakh	-да, -де, -та, -те (-da, -de, -ta, -te)	-ға, -ге, -қа, -ке (-gha, -ge, -qa, -ke)
Kirghiz	-да, -де, -до, -дө, -та, -те, -то, -тө (-da, -de, -do, -dö, -ta, -te, -to, -tö)	-га, -ге, -го, -гө, -ка, -ке, -ко, -кө (-ga, -ge, -go, -gö, -ka, -ke, -ko, -kö)
Tatar	-да, -дә, -та, -тә -da, -də, -ta, -tə	-га, -гә, -ка, -кә -ga, -gə, -ka, -kə
Turc	-da, -de, -ta, -te	-a, -e, -ya, -ye
Turkmène	-да, -де -da, -de	-а, -ә -a, -e
Ouïghour	دا، -ده، -تا، -ته (-da, -de, -ta, -te)	غا، -گه، -قا، -كه، -قه، -كا (-gha, -ge, -qa, -ke, -qe, -ka)

Ouzbek	-да	-га
	-da	-ga
Français	dans (prep.)	à (prep.)

Tableau 2. Une comparaison des prépositions. Les translittérations ou équivalences en écriture latine apparaissent dans la 2^{ème} ligne.

A partir de cette base, nous pouvons former des nouveaux mots en rattachant des morphèmes. Par exemple:

Kazakh	turc	ouïghour	traduction
дин + да (din+da)	din+da	(din+da) دين+دا	dans la religion
N. + LOC			
Turc	ouïghour	ouzbek	traduction
gök + e	كۆك+ه (kök+ke)	ko'k + ga	à bleu, au ciel
N. + DAT.			

Tableau 3. Formation d'un nouveau mot en plusieurs langues

On remarque que dans le premier exemple que les trois langues utilisent le même suffixe « da ». Par contre, dans le deuxième exemple chacune des langues utilisent des suffixes variés. Cette variation, naturellement, multiplie la différence de la prononciation.

Etant aussi des langues agglutinantes et les autres langues turques partagent les caractéristiques de la langue ouïghoure que nous avons mentionnées dans le paragraphe 2.2.5. Mais il existe bien des différences entre chaque langue en raison de la diversité de l'écriture, des sons et de l'inflexion des affixes. Par conséquent, la problématique de la morphologie demande une étude particulière pour toutes les langues. Dans le prochain chapitre, nous ferons une étude plus approfondie de la morphologie de l'ouïghour.

Chapitre 2. Une introduction à l'écriture ouïghoure-latine

Dans ce chapitre nous révisons l'histoire de l'échec de la réforme visant à substituer à l'écriture ouïghoure basée sur l'alphabet arabe par un alphabet latin inspiré par le Pinyin, et nous présentons les études menées dans le cadre de proposition d'un nouveau standard de transcription – l'écriture latine ouïghoure (LSU). Nous exposons les règles d'orthographe qui s'y appliquent, le cadre actuel de son utilisation et ses perspectives d'avenir.

2.1 Introduction

Les documents historiques nous montrent que les Ouïghours ont utilisé divers alphabets, parfois en concurrence les uns avec les autres, durant différentes époques au cours des dernières 2000 années. Les plus importants sont les alphabets soghdien (5^e-10^e siècle⁵³, puis 15^e-16^e siècle⁵⁴), Orkhon-Yéniseï, aussi connu sous le nom d'alphabet turco-runique, (6^e-9^e siècle)⁵⁵, ouïghour ancien (10^e-18^e siècle)⁵⁶ puis, après l'adoption progressive de l'Islam à partir de 934 de notre ère, l'alphabet arabe modifié (10^e siècle à aujourd'hui).

Modifié afin de prendre en compte les caractéristiques phonétiques et grammaticales des langues turques, l'alphabet arabe (parfois appelé l'alphabet *Chaghatay*) a été utilisé continuellement par les Ouïghours et les autres peuples turcophones d'Asie centrale jusqu'aux années 20. Plusieurs tentatives de réforme ont ensuite été lancées. Les changements d'orthographe visant à rationaliser la graphie des mots d'origine turque et à assimiler les mots d'origine arabe et persane ont petit à petit mené à diverses tentatives de passage aux alphabets latin et cyrillique. Beaucoup de ces projets avaient une motivation autre que linguistique : pour la plupart, ils étaient directement liés à la « politique des nationalités » de Staline, établie par les décisions du premier Congrès International de

⁵³ Haji Nurhaji, *Qedimki Uyghurlar we Qaraxaniylar* (Ouïghours anciens et Qarakhanides), Xinjiang People's Press, 2001, p.122.

⁵⁴ Abdurêhim Rahman & Ömerjan, *Shinjang Tarix Matériyalliri* (Documents Historiques sur le Xinjiang), volume 40, Xinjiang People's Press, 1980.9.

⁵⁵ On croit que l'alphabet Orkhon est un dérivé d'un alphabet soghdien non-cursif. Il a été utilisé lors de l'empire Orkhon ouïghour (744-840) mais aussi par d'autres tribus turques de l'Asie centrale occidentale. Voir <http://www.omniglot.com/writing/orkhon.htm>

⁵⁶ Adopté, en forme modifiée, par les Mongols puis par les Mandchous.

Turcologie qui s'est tenu à Bakou en 1929 et avait pour but de romaniser les « langues littéraires » récemment créées à pour chacune des langues turques d'Asie centrale. L'adoption par la République de Turquie de l'alphabet latin entre 1927 et 1928 a contribué à influencer la décision en 1937 de l'Union Soviétique, concerné par une menace « panturque », d'abandonner l'alphabet latin et d'imposer des alphabets basés sur le cyrillique à ses peuples turcs⁵⁷.

La Chine, où vivent la majorité des Ouïghours, a été particulièrement active dans les réformes de langues et d'alphabets après la victoire du parti communiste en 1949. Elle a dès lors repris à son compte l'approche soviétique aux politiques linguistiques. A cet époque, les linguistes russes ont vivement aidé la Chine à répertorier et codifier les langues des minorités ethniques (les Chinois non-Han) vivant sur son territoire. Dans ce contexte, les dialectes ouïghours, parlés des deux côtés de la frontière entre l'URSS et la Chine, ont été unifiés en une seule « langue littéraire » basée sur les normes utilisées pour les Ouïghours vivant en URSS. Dans la foulée, en mars 1956, le Conseil d'Etat de la République Populaire de Chine (RPC) et le Comité⁵⁸ des Langues et des Ecritures de la RAOX⁵⁹ (*Shinjang Uyghur Aptonom Rayonluq Til-Yéziq Komitétí*, ci-après XLSC) ont lancé une campagne de courte durée afin d'imposer l'alphabet cyrillique⁶⁰ aux Ouïghours, Kazakhs et Kirghiz (tous les trois turcophones) mais aussi aux Mongols et Sibes (*xībó*) qui vivaient sur son territoire.

Lorsque ses relations avec l'URSS sont entrées dans une période de forte tension, la Chine a décidé de rebrousser chemin sur les politiques linguistiques qui créaient trop de liens entre les Ouïghours (et autres peuples turcophones d'Asie Centrale) qui vivaient de part et d'autre de la frontière sino-soviétique. En novembre 1959, la Chine a promulgué un projet de proposition d'un alphabet latin basé sur le « *Hanyu Lading-hua Pinyin* » (*latinisation du Chinois*) appelé en ouïghour *yéngi yéziq* ou « nouvelle écriture » pour

⁵⁷ Voir Kamal Talibzade, *Learning to Read All Over Again: Alphabet Changes in Azerbaijan Throughout the Century, An Eyewitness Account*, Azerbaijan International 2000 (8.1),

http://www.azer.com/aiweb/categories/magazine/81_folder/81_articles/81_talibzade.html

⁵⁸ Une filiale gouvernementale créée avec le but de « promouvoir le développement des langues minoritaires (chinois non-Han) et systèmes d'écriture dans la RAOX ».

⁵⁹ Région Autonome Ouïghoure du Xinjiang

⁶⁰ Cet alphabet modifié du cyrillique était, pour le ouïghour, emprunté directement par celui qui avait été imposé aux ouïghours de l'union soviétique en 1947 afin de remplacer l'alphabet romanisé d'origine décidé en 1928.

l'ouïghour (ci-après PNSU), qui a été adopté⁶¹ pour l'ouïghour et le kazakh à la première Conférence Linguistique et Orthographique Nationale de la RAOX et approuvé par le Congrès du Peuple de la RAOX la même année. Entre 1960 et 1964, le PNSU a été mis en œuvre à titre de projet pilote, puis promulgué dans l'ensemble de la RAOX. Cet alphabet est alors devenu l'unique alphabet autorisé pour l'ouïghour et le kazakh jusqu'à ce que l'alphabet arabe (ci-après ASU⁶²) soit officiellement ré-autorisé en septembre 1982 (document n. XH-1982-283 du gouvernement de la RAOX), après avoir subi de nouvelles modifications de détail. Même après le retour en grâce de l'ASU et sa réadoption enthousiaste par la population, le PNSU a conservé un statut d'option officielle autorisée⁶³.

2.2 Historique

La création et la promulgation de PNSU doivent aussi être examinées en relation avec l'environnement linguistique général de la RPC à cet époque. Dans un effort de lutte générale contre l'analphabétisme, le pouvoir de Pékin avait décidé que les caractères chinois devraient être remplacés à terme par l'alphabet latin (*Hanyu pinyin*, ou *pinyin* simplement) et la réforme de l'écriture des langues ethniques « minoritaires » devait s'intégrer dans un plan général qui prévoyait un alphabet commun pour tous les groupes ethniques de la Chine.

L'échec essuyé en fin de compte par cette tentative de romaniser l'écriture ouïghoure (ainsi que d'autres systèmes d'écriture de la RAOX) peut être attribué aux facteurs suivants :

La résistance culturelle : L'alphabet arabe représentait le lien avec la tradition religieuse (l'Islam) et le passé culturel pour la grande majorité des Ouïghours attachés à leur identité et ils rejetaient la réforme imposée par les autorités comme un pas de plus vers l'assimilation et la perte de leur autonomie linguistique et culturelle.

⁶¹ Voir “*wéiwúěr xīn wénzì fāng'àn — cǎo'àn*” (document de travail pour la nouvelle écriture ouïghoure — préliminaire) 《维吾尔新文字方案》(草案) and “*hāsàkè xīn wénzì fāng'àn — cǎo'àn*” (document de travail pour la nouvelle écriture kazakh — préliminaire) 《哈萨克新文字方案》(草案), novembre 1959,

http://www.xjyw.gov.cn/han/wenzi_gai.htm

⁶² Voir <http://www.omniglot.com/writing/uyghur.htm>

⁶³ Sur la réforme d'écriture et de langue, voir J. R. Duval: “Modern Uyghur, A Historical Perspective” dans: *Culture Contact, History and Ethnicity in Inner Asia*, № 2. Michael Gervers and Wayne Schlepp, eds. Toronto: Joint Center for Asia Pacific Studies, 1996, pp. 132-67.

Un sentiment d'éloignement des autres peuples turcs : Les réformes de l'écriture en Turquie et dans les autres communautés turcophones qui avaient créé de multiples transcriptions latines et cyrilliques pour des sons similaires. La Chine, dans son effort de romanisation, était allé un pas plus loin en introduisant des nouvelles lettres uniques (Θ , Օ , Ө , Ƶ , Ƙ et Һ⁶⁴), qui isolait l'ouïghour des autres langues turques romanisées. L'utilisation, adoptée du *pinyin* chinois, des lettres *q* et *x* pour les phonèmes [tʃ] et [ʃ] ne faisaient que renforcer ce sentiment⁶⁵.

La peur de l'assimilation linguistique : L'introduction des groupes de lettres *ch*, *sh* et *zh*, utilisés afin de transcrire des phonèmes purement chinois, ont rendu possible l'importation de mots chinois directement dans la langue ouïghoure. Ceci a été vu comme une autre manière d'accélérer le processus de sinisation⁶⁶ de la langue. Ce qui est intéressant est que PNSU n'allait pas jusqu'au bout en imposant l'orthographe chinoise *pinyin*, puisque les mots chinois importés en ouïghour devaient quand même subir certains changements, telle la substitution du *y* au *i* du *pinyin* dans les diphtongues *ai* et *ei*, et du *a* au *e* ou *a* du *pinyin* dans certains environnements phonétiques, sapant ainsi le but de l'unification linguistique.

La fatigue des réformes : En très peu de temps, l'ouïghour a subi toute une série de réformes alphabétiques (des modifications apportées à l'arabe au changement total au cyrillique puis l'alphabet latin). Entre temps, l'ASU a aussi subi des modifications qui l'ont

⁶⁴ Chaque Θ, Ө, a été emprunté de l'alphabet cyrillique et modifié pour les langues d'Asie centrale. Ƶ, Һ et Օ sont des créations de la RPC.

⁶⁵ Pour les lettres *q* et *x*, et la lettre *j*, il est intéressant de constater qu'elles ne transcrivent pas en PSNU les mêmes phonèmes qu'en *pinyin*, où elles correspondent respectivement à [tʃ^h], [ç] et [tʃ], et non à [tʃ] [ʃ] et [dʒ]. Pour les Ouïghour non sinophones, il n'y a aucune différence de prononciation entre ces phonèmes. Voir <http://en.wikipedia.org/wiki/Pinyin>

⁶⁶ Beaucoup de nouveaux mots, surtout de nature politique, ont été importés directement du chinois au ouïghour durant cette époque, qui correspond à la Grande Révolution Culturelle Prolétarienne et à une série de campagnes contre le nationalisme ouïghour. La plupart de ces mots ont disparu, seulement pour être remplacés par des termes « natifs » lorsque l'alphabet arabe a été réintégré (voir J. R. Duval, *op.cit.* note 67 ci-haute).

rendu beaucoup plus simple d'utilisation pour une langue non-arabe⁶⁷, réduisant ainsi la nécessité de le remplacer.

Puisque des millions d'ouïghours de la RAOX ont été éduqués pendant presque vingt ans (entre 1965 et 1982) en PSNU et non en ASU, ils sont devenus analphabètes lorsque l'ASU est réapparu en tant qu'alphabet officiel. En même temps le CSU⁶⁸ était et est toujours l'unique alphabet utilisé par les ouïghours⁶⁹ soviétiques.

Cette situation chaotique a créé des obstacles pour les Ouïghours, qui ne pouvaient communiquer par écrit entre eux du fait de l'existence de frontières géopolitiques mais aussi intergénérationnelles. L'arrivée de nouvelles technologies de communication a malheureusement amplifié le problème au lieu de le résoudre.

Au cours des deux dernières décennies, les ordinateurs et l'équipement de technologie moderne ont fait naître de nouveaux espoirs dans le domaine du traitement d'information et de la communication. Malheureusement, un manque d'unification des alphabets et l'existence d'une variété de caractères uniques en PNSU, CSU et ASU, ont ralenti l'arrivée du progrès des technologies d'information. Le manque de standards complets pour ASU, CSU et PNSU, autant dans l'ASCII que dans l'Unicode, ont forcé les utilisateurs ouïghours de recourir à une grande variété de translittérations non-normalisés afin de pouvoir partager et traiter l'information dans leur langue native (ou à d'autres langues – chinois, russe et anglais, majoritairement). D'un côté, cela a permis aux internautes ouïghours de se familiariser avec l'alphabet latin et a accéléré la mort du PNSU,

⁶⁷ Les lettres arabes utilisées purement pour des phonèmes arabes (ث, ذ, ح, ط, ظ, ص, ض, ع) ont été abandonnées petit à petit à partir des années 20, tandis que d'autres (notamment و, ه, ا, ي) ont été investies d'un nouveau rôle, celui d'indiquer des voyelles—avec, en position initiale l'adjonction du *hamza sur une dent* afin de remplacer le *alif* silencieux initial traditionnel de l'alphabet arabe. Parmi les créations de nouvelles lettres, notons celle d'une **ه** [e] initiale; la spécialisation progressive de **و**, premièrement avec **و** [o, ɔ] puis **و** [u, ü], et plus tard avec **و** [o], **و** [ɔ], **و** [u] et **و** [ü]; et la création de la lettre **ي** [i] pour marquer le i par opposition au **ي**, plus tard supplémenté par **ي** [é] afin de transcrire les sons de voyelles absents de l'alphabet traditionnel. A travers ces modifications ASU est devenu un alphabet phonémique avec une lettre pour chaque son distinct. En même temps, il est devenu un alphabet indépendant qui ne suivait plus les règles traditionnelles de l'alphabet arabe.

⁶⁸ <http://www.omniglot.com/writing/uyghur.htm>

⁶⁹ Voir annexe 10

vu que ce dernier comportait des caractères qui sont absents de la plupart des systèmes informatiques. D'un autre côté, cela a causé la multiplication de systèmes romanisés, mutuellement incompatibles sur les plateformes d'information.

En 2002, d'après un sondage réalisé par l'auteur, il y avait au moins dix-huit systèmes de romanisation activement utilisés pour l'ouïghour. Ce nombre n'inclut pas les translittérations non systématiques qui apparaissaient ici et là en raison de l'absence d'un système commun. Certains ont un nom (Angel, Kiboon, Makanim, ML, Hashim, PBLSU, UIY [*écriture ouïghoure surl'internet*], Zaman) et certains n'en ont pas. Des érudits ont proposé différents alphabets influencés par le pays dans lequel ils vivaient ou de la langue étrangère parlée. Il est facile de détecter, parmi les dix-huit alphabets, des influences turques, allemandes, russes, anglaises et chinoises, ainsi que divers éléments créatifs imposés sur les alphabets.

Entre novembre 2000 et juillet 2001, cinq conférences⁷⁰ se sont tenues à l'Université du Xinjiang à Ürümchi afin d'introduire un alphabet ouïghour unifié (ci-après LSU⁷¹) utilisant les caractères latins. Parmi ceux qui assistaient à la conférence, il y avait des linguistes ouïghours tels Ibrahim Mutihi, Mirsultan Osman, Muhebbet Qasim, Imin Tursun, Abliz Yaqup, Xemit Zakir, Arslan Abdulla, ainsi que des experts en informatique tels Hoshur Islam, Türgün Ibrahim, l'auteur de cette thèse⁷² et des jeunes étudiants. Pour des raisons politiques et culturelles évidentes, les conférences ont fortement insisté sur le fait que « l'alphabet proposé ne devrait pas remplacer l'ASU, et son introduction ne devra pas représenter une réforme de l'alphabet actuel. Son utilisation sera uniquement dans les domaines de l'informatique comme système d'écriture accessoire. »

Afin d'atténuer les craintes d'encore une nouvelle réforme de l'alphabet, et souligner la portée limitée du projet, il a été décidé d'appeler ce système de romanisation « *Uyghur Kompyutér Yéziqi* » (UKY) ou écriture ouïghoure informatique. Ce nom a été changé par

⁷⁰ Voir http://www.ukij.org/teshwiq/UKY_Heqqide.htm pour plus de détails.

⁷¹ A l'origine appelé, en ouïghour, *Uyghur Kompyutér Yéziqi* (UKY)—*écriture ouïghour en informatique*, puis *Uyghur Latin Yéziqi* (ULY)—*écriture ouïghoure latine*. Voir [www.ukij.org/teshwiq/UKY_Heqqide\(KonaYeziq\).htm](http://www.ukij.org/teshwiq/UKY_Heqqide(KonaYeziq).htm)

⁷² Qui a initié toutes ces conférences et préparé des propositions d'étude et de recherche qui sont devenues des documents de travail de base pour l'unification de l'LSU.

l'Association Informatique Ouïghoure (*Uyghur Kompyutér Ilimi Jem'iyiti – UKIJ*⁷³) trois ans plus tard pour devenir *Uyghur Latin Yéziqi* (ULY, écriture ouïghoure latine), ou *Latin Script Uyghur* (LSU) puisque en ouïghour « écriture informatique » était mal perçu et pouvait être confondu avec le binaire – 1 et 0.

Le projet LSU a été couvert par les médias officiels de la RAOX ainsi que sur l'internet afin d'aviser le public. Les médias en particuliers étaient très attentifs à ne pas donner l'impression d'un début de réforme de l'écriture. N'empêche que même aujourd'hui, certains hésitent à utiliser le terme ULY (LSU) puisqu'ils craignent l'association potentielle à une tentative de réforme de l'alphabet commun. Pour beaucoup, il est important qu'il y ait une correspondance lettre pour lettre (ou une norme) entre LSU et ASU. Certains estiment aussi que l'ouïghour devrait à l'instar du turc, de l'azerbaïdjanais et de l'ouzbek adopter exclusivement l'alphabet latin (des plans similaires ont été faits, mais toujours pas implémentés dans d'autres républiques turques, tels le Kazakhstan, le Kirghizstan et le Turkménistan), mais cela va à l'encontre des politiques officielles d'état.

Un rapport a été envoyé aux autorités gouvernementales de la RAOX au printemps 2001 et, en guise de réponse, le comité du peuple RAOX a organisé une audition qui a reconnu la valeur du projet et recommandé que des tests à grande échelle soient effectués afin d'évaluer sa faisabilité.

2.3 Principes de base utilisés dans la création du LSU

Après multiples réformes, l'alphabet ASU avait évolué de manière significative. Il transcrit maintenant lettre par lettre tous les phonèmes de la langue ouïghoure, même si ce n'est pas une transcription phonétique exacte, puisqu'il reflète les sons théorétiques de base et non les changements multiples qui se produisent dans la langue ouïghoure parlée⁷⁴. Depuis 2001, un certain nombre de règles d'orthographe ont été promulguées par la XLSC⁷⁵ afin de

⁷³ Une association à but non lucratif fondée en janvier 2004 par l'auteur de cette thèse. Site web: <http://www.ukij.org>.

⁷⁴ Telle la prononciation dans la plupart des dialectes de *n* comme *m* devant les labiales, l'adoucissement de *q* à [x] devant une autre consonne, de *j* à [ʒ] devant une autre consonne, l'adoucissement des lettres finales de racine *j* (→[ʒ]), *k* (→[g]), *p* (→[b]/[w]), *q* (→[ɣ]), et *t* (→[d]) devant un suffixe commençant par une voyelle, l'abandon fréquent des lettres *r* et *l* et dans une position post-vocalique ou le changement de *l* en *m* devant un *m*.

⁷⁵ L'orthographe officielle est établie par XLSC. Etant donné le nombre de réformes partielles qu'a subies l'orthographe officielle au cours des dernières années et l'impact des prononciations dialectales, un grand nombre de

codifier son utilisation. Malgré des réformes successives, les règles d'orthographe contiennent toujours un nombre d'incohérences, surtout avec les lettres finales *b/p* et *d/t* dans des mots d'origine étrangère (généralement arabe et persane mais aussi russe), lorsque l'orthographe officielle hésite entre l'étymologie et la vraie prononciation⁷⁶.

Le LSU est une transcription de la langue officielle des Ouïghours de la RAOX et suit les mêmes règles d'orthographe que l'ASU.

Dans cette partie, basée sur la participation de l'auteur aux travaux du comité et sur la documentation de ses débats et décisions, nous décrivons les quatre principes de base qui ont guidé le travail du XLSC et qui ont donc donné lieu aux décisions prises afin de créer le LSU.

A. Correspondance une-à-une entre les lettres de LSU et les phonèmes de la langue écrite qui donne donc, au mieux possible, une cohérence entre les caractères de LSU, ASU et CSU⁷⁷.

L'objectif était de faciliter l'apprentissage du LSU pour des personnes utilisant des alphabets différents en évitant l'obstacle potentiel de nouvelles règles de transcription/orthographe.

B. Absence d'ambiguïté et éviter au plus possible de doubler les lettres.

Ce principe était afin d'encourager la création d'un système de transcription/translittération claire et logique qui permettrait une lecture facile, qui empêcherait des ambiguïtés possibles (particulièrement dans le cas des voyelles combinées qui créent un seul son) et d'éviter l'allongement inutile des mots.

variantes peut être retrouvé dans les documents écrits de la RAOX. Dans le cas du CSU, l'orthographe est généralement basée sur les dialectes ouïghours du nord (essentiellement caractérisés par une substitution fréquente du son [ʒ] en position initiale pour *y*) et a tendance à suivre la vraie prononciation, par exemple en abandonnant le *t* final non prononcé après *s* ou *sh* mais en le réinsérant lorsqu'un suffixe exige sa prononciation.

⁷⁶ Ex. Trois mots venant de la même racine trilitère arabe k-t-b, كتاب (*kitab*—livre), كۈتۈپخانا (*kutupxana*—bibliothèque) et مەكتەپ (*mektep*—école); ئىقتىساد (*iqtisad*—économie, prononcé *ixtisat*, de l'arabe اقتصاد *iqtisad*) versus مەقسەت (*meqset*—but, prononcé *mexset*, de l'arabe مقصد *maqsad*). Voir *Hazirqi zaman Uyghur edebiy tilining teleppuz lughiti* (Dictionnaire de la Prononciation de l'Ouïghour Moderne), Nationalities Publishing House, Beijing 1988).

⁷⁷ Pour les lettres CSU, la correspondance une-à-une ne peut être garantie que dans des textes suivant l'orthographe de la RAOX.

C. Utilisation minimale de signes diacritiques, lesquels devraient être restreints aux plus communs.

A cette époque, la plupart des claviers présents dans la RAOX étaient du type QWERTY nord-américain et peu d'utilisateurs étaient susceptibles d'installer des logiciels spécifiques pour LSU, les ordinateurs étant programmés par défaut pour l'entrée de données en anglais ou en chinois.

D. Une correspondance approximative des lettres latines choisies avec leur valeur phonétique internationale.

Ce principe avait pour but d'assurer que les lettres puissent être utilisées de manière consistante avec la prononciation internationale commune et ce, en partie, afin d'éviter la confusion parmi les non-ouïghourophones qui tenteraient de lire, par exemple, un prénom ou nom géographique en ouïghour (quel lecteur de langue anglaise, française, allemande ou espagnole saurait comment prononcer approximativement Üürümchi si celle-ci était écrite *Üürümqi* ou *Vurvumqi*, ou *Shemshi* écrite *Xəmxı*).

Les principes énumérés ci-dessus ont été appliqués de manière hiérarchique, avec la priorité accordée au premier d'entre eux, tout en tenant compte de l'importance des principes qui suivent. L'alphabet PNSU des années soixante et soixante-dix n'a pu passer le test d'aucun de ces principes: il y avait plus de lettres et de groupes de lettres que dans l'ASU, il était ambigu puisqu'il proposait deux transcriptions pour certains sons (*q* et *ch*, *x* et *sh*, *j* et *zh*) ; il utilisait des signes diacritiques et des lettres absents de l'alphabet latin commun, rendant les mots ouïghours illisibles pour tous ceux qui n'étaient pas familiers avec le PNSU, imposant donc un recours à d'autres types de transcription des prénoms et noms géographiques à l'extérieur de la RAOX. Sur les cartes romanisées de Chine, la plupart des noms de villes et de lieux ouïghours apparaissent avec leur nom *pinyin* – *Yining* pour *Ghulja* et *Kashi* pour *Qeshqer*. Les noms et prénoms ouïghours apparaissent sur les documents d'identité avec la transcription *pinyin* de leur translittération en caractères chinois - *Waresijiang Abudukelimu* au lieu de *Warisjan Abdukérım*. Il a cependant été difficile pour ceux qui avaient appris le système PNSU d'apprendre une nouvelle transcription et une nouvelle valeur phonétique. Le statut semi officiel retenu par le PNSU aurait pu décourager l'apprentissage du LSU mais l'exemple d'autres nations turcophones d'Asie centrale qui

adoptaient à l'époque l'alphabet latin a aidé à faire progresser rapidement le projet. Sur la base des principes énumérés ci-dessus, une norme provisoire a finalement été adoptée par les participants⁷⁸ de la conférence. Plus bas, nous décrivons les détails des décisions prises pour chaque lettre.

2.4 Lettres ne posant pas de problème particulier

Parmi la variété des transcriptions en utilisation, il était possible d'identifier les dix-huit lettres suivantes qui étaient employées par tous pour les mêmes phonèmes. Puisqu'elles correspondaient aux principes ci-dessus, elles ont été acceptées sans problème dans l'alphabet LSU par le comité de la conférence.

م	ل	ك	ئي	گ	ف	د	ب	قا
m	l	k	l	g	f	d	b	a
ز	ي	ئو	ت	س	ر	پ	ئو	ن
z	y	u	T	s	r	p	o	n

Table 1 Lettres ne posant pas de problème particulier

2.5 Lettres problématiques et décisions

Pour les autres lettres il existait les choix suivants, basées sur les systèmes de transcription variés :

18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	
q, ch	q	Q	q	Q	č	č	ç	ç	ç	ç	c	ch	ch	ch	ch	ch	ch	چ
ə	ae	Va	e	E	ä	ä	ä	é	e	e	e	ä	e	a'	e	e	e	ئه
E	e	E	e	E	e	e	e	e	é	é	e	e	e	e	ë	e	i	ئي
□	gh	Vg	w	V	ğ	ğ	ğ	ğ	ğ	ğ	gh	gh	gh	gh	gh	gh	gh	غ
h	v	Vh	wh	Hf	h	h	h	h	h	h	h	h	h	h	h	h	h	ه
j, zh	j	J	j	J	j	d ₃	j	j	c	c	j	j	j	j	j	c	j	ج
Ng	ng	Ng	nh	Nf	ŋ	ŋ	ñ	ŋ	ñ	ñ	ng	ng	ng	ng	ng	ng	ng	ڭ
ə	oe	Vo	oe	Oi	ö	ö	ö	o'	ö	ö	o	ö	o	o'	ö	v	o	ئو
K	kh	Vk	k	C	q	q	q	k	q	k	q	q	q	q	q	q	q	ق
x, sh	x	X	x	X	š	š	ş	ş	ş	ş	x	sh	sh	sh	sh	sh	sh	ش
Ü	ue	Vu	ue	Ui	ü	ü	ü	ur	ü	ü	u	ü	u	u'	ü	y	u	ئو
W	w	W	v	W	w	v	v	v	v	v	w	w	v	v	v	w	v	ؤ

⁷⁸ Voir http://www.ukij.org/teshwiq/UKY_Heqqide.htm pour plus de détails.

H	h	H	h	H	x	x	x	x	h̃	h	h	kh	x	x	x	x	x	ç
z	zj	Vz	zh	Zf	ž	з	ž	ž	ž	ž	jz	zh	j	zh	zh	zx	j	ž

Table 2. Lettres problématiques et leurs équivalences latines proposées

Il y a fallu faire des choix selon les principes de base. Les discussions suivent :

چ [tʃ]: Le principe C a éliminé ç et ċ et le principe D a enlevé c et q. **Ch** a été choisi car il respectait toutes les règles, même s’il avait l’inconvénient d’être la combinaison de deux lettres. Il se lit [tʃ] dans la plupart des alphabets basés sur le latin dans le monde (à l’opposé du PNSU q, ç turc, ċ slave ou, plus simplement, c, proposé dans quelques alphabets) et a été choisi pour le même phonème dans l’ouzbek latin. Il est facilement lisible pour les ouïghours éduqués en PNSU puisqu’il était utilisé dans cet alphabet pour le phonème [tʃ] dans les mots d’origine chinoise (ex. *gungchəndang* – parti communiste).

ئ [æ]: le principe B a éliminé ə et le principe D a éliminé va. Les lettres combinées ae avaient le désavantage de rallonger des mots inutilement comme ئ est une voyelle fréquemment utilisée (ex. *maektaep* pour مەكتەپ *mektep* – école). Le même a été dit de l’utilisation de l’apostrophe après un a qui a été proposé afin de suivre l’exemple de l’ouzbek latinisé afin d’indiquer une lecture variante des lettres latines g et o (école devrait alors s’écrire *ma’kta’p*). D’ailleurs, l’utilisation de l’apostrophe aurait pu aboutir à des ambiguïtés, comme nous allons le voir après. Un certain nombre de personnes était en faveur d’utiliser la lettre e. Mais, comme [æ] est l’une des trois voyelles antérieures de l’ouïghour, beaucoup pensaient qu’il serait mieux d’utiliser ā pour le transcrire, afin d’avoir une opposition visuelle systématique entre les voyelles postérieures a/o/u et les voyelles antérieures ā/ō/ü, comme dans les cas du hongrois et du finnois. Ils pensaient que e devrait être utilisé pour ئې, déjà transcrit comme e en PNSU. Ce sont des considérations de fréquence⁷⁹ qui ont mené à l’adoption finale de la solution du turc, e.

⁷⁹ D’après Alim Ahat de la compagnie Uighursoft (www.uighursoft.com) ئە apparaît 67,121 fois (4.065%) dans un corps de dictionnaire de 1, 651, 130 lettres, alors que ئې apparaît 13,037 fois (0.790%).

ئې [e]: Puisque *e* n'était plus disponible (voir ci-dessus), le choix devait se faire entre *i*, *ë* et *é*. Le principe A a éliminé *i*. En fin de compte, *é* a été choisi puisque l'accent était d'une utilisation plus aisée que le tréma (une seule marque diacritique au lieu de deux). De toute façon, le son est plus proche du son du *é* français que du *ë* albanais.

غ [ɣ]: Les principes B,C et D ont éliminé *ğ*, *g*, *v*, *vg*, *w* et *oɣ*, et la lettre *g* était déjà occupée par گ. La lettre double **gh** a été choisie par opposition au *g'* ouzbèke puisqu'elle est d'usage commun dans la représentation du غ dans les mots venant de langues où existe le son [ɣ] : ex. Af**gh**anistan, Mag**h**reb, A**gh**a Khan.

ھ [h]: Les principes B, C et D ont éliminé *v*, *h*, *vh*, *wh* et *hf*. La lettre **h** a été choisie comme étant la solution la plus simple (sans marque diacritique) et aussi parce que, dans la plupart des langues, elle représente un son très proche du son ouïghour.

چ [dʒ] : Le principe B a éliminé *ɟʒ*, le principe D a éliminé *c* et *zh*. La lettre *c* est prononcée [dʒ] en turque et azéri. Pour ce son, PNSU offrait deux choix, : *j* dans les mots natifs et *zh* dans les mots d'origine chinoise (ex. *Fangzhen*), ce dernier étant utilisé afin de représenter ce son uniquement dans le *pinyin* chinois. **J** a été choisi comme étant plus proche de l'usage international.

ڭ [ŋ]: Le principe B a éliminé *ñ*, *ñ* et *ŋ*, et le principe D a éliminé *nh* et *nf*. Le choix a naturellement fini par être la lettre composée **ng**, qui est utilisé communément dans beaucoup de langues afin de représenter le son [ŋ].

ئۆ [ø]: Les options *vo*, *oe* et *oi* ont été éliminées puisqu'elles allongeaient inutilement les mots ou ne rendaient pas les sons ouïghours assez clairement pour les non ouïghourophones. La lettre *v*, choisie par certains pour sa similitude avec la partie de haut de la lettre ASU ئۆ a été éliminée sur la base du principe D. La lettre PNSU **ø** a été éliminée sur la base du principe B. Etant donné le besoin de le différencier avec *o*, il a été décidé d'utiliser la lettre **ö** qui, malgré l'inconvénient d'un signe diacritique, est utilisée communément dans d'autres langues afin de transcrire un son similaire ou proche de ce son en ouïghour.

ق [q]: Les principes B et D ont éliminé *k* et *κ*. Le principe D a enlevé *vk* et *c*. La lettre *k* était déjà utilisée pour ك et la lettre composée *kh* correspond à un phonème différent dans tous les systèmes internationaux. La lettre **q** a été choisie comme elle est la transcription commune utilisée pour la même lettre dans les alphabets basés sur l'arabe(ex : al-Qaida, Mohammed Iqbal, Iraq, Qatar). C'est aussi la lettre utilisée pour le même son dans l'ouzbek latin et le tatar, et pour un son connexe en azéri.

ش [ʃ]: Les principes B et D ont éliminé *ş* et *š*. Le principe D a éliminé *x* qui est utilisé dans seulement quelques systèmes graphiques (majoritairement *pinyin*, portugais et vietnamien) afin de transcrire ce phonème. L'usage international ainsi que l'ouzbek et l'azéri latinisés ont dicté la décision de choisir **sh**. Cette décision a été renforcée par le fait que, à part le *x*, le PNSU propose aussi *sh* pour des mots d'origine chinoise telle *gungshe* (commune)⁸⁰.

ئو [y]: Les principes C et D ont éliminé *u'*, *vu*, *ue* et *ui* et le principe A a éliminé *y* (un choix attractif car c'est le symbole IPA pour la voyelle haute antérieure arrondie) puisqu'elle était déjà utilisée pour ي, et *u* qui représente ئو. La lettre **ü** a finalement été choisie puisqu'elle utilise la même méthode de différentiation avec son équivalence vocalique postérieure *u* que *ö* vis à vis de *o*⁸¹, et est fréquemment utilisée dans le monde pour représenter un son similaire ou proche du son ouïghour ئو.

ۋ [w]: Cette lettre est utilisée afin de rendre une consonne qui alterne entre la bilabiale [w] et la fricative labio-dentale [v] : un son consonantique qui rappelle la lettre néerlandaise *w* (quelque par entre *w* et *v*), en début de syllabe (ex. ۋەتەن [wætæn]—patrie, ۋەزىر [wæzir]—vizir, قولۋاق [qolwaq]—chaloupe) ; ainsi qu'une semi-consonne bilabiale lâche [ɰ] en fin de syllabe ou après une voyelle (ex. لەۋ [ləw]—lip). Etant donnée cette double valeur

⁸⁰ En *pinyin*, la lettre composée *sh* représente la fricative post-alvéolaire sourde [ʃ] tandis que la lettre *x* représente la fricative alvéolo palatale sourde [ç]. En ouïghour, seul le premier de ces deux sons existe. Dans l'ASU, le *sh* et *x* du PNSU sont transcrits par la lettre ش.

⁸¹ Il faut noter qu'en PNSU, les voyelles postérieures et antérieures ne sont pas différenciées de manière systématique : aux voyelles postérieures *a*, *o* et *u*, correspondent dans le registre des voyelles antérieures une nouvelle lettre, *ə*, une lettre avec une marque diacritique intérieure, *ø*, et une lettre avec une marque diacritique extérieure, *ü*.

phonétique, beaucoup de pensaient utiliser la lettre *v* en position initiale et *w* en position finale. Ceci allait, bien sûr, contre le principe A au niveau de la correspondance entre LSU, ASU et CSU, et il a été décidé de ne choisir qu'une seule lettre. La décision finale a favorisé *w*, en partie puisqu'elle se lit mieux en position finale que *v*.

خ [x]: Il était difficile de trouver une lettre latine idéale afin de représenter خ. Beaucoup, habitués au PNSU, croyaient que la lettre *h* serait le meilleur choix puisqu'elle correspondait à l'usage du *pinyin* chinois. Un autre argument en sa faveur était qu'elle éviterait un recours à une autre lettre composée, *kh*, communément utilisé en Occident pour la transcription de langues basées sur l'arabe ou le cyrillique. Mais *h* était utilisé par une autre lettre (ه) et, dans la plupart des langues, n'est pas utilisé afin de représenter le son خ. Certains insistaient qu'en l'absence d'une solution satisfaisante, la lettre *h* devrait être utilisée pour les deux sons et que le contexte aiderait à les différencier⁸². Selon le principe A, cela était inacceptable. Le choix final est donc tombé sur *x*, une lettre communément utilisé dans des transcriptions à caractère scientifique et adoptée par l'azerbaïdjanais et l'ouzbek latinisés. Cette solution « du moindre mal » ne correspond pas totalement au critère D. Ainsi, le nom *Xoten* d'une ville importante du sud du Tarim pouvant être prononcé [zouten] par les anglophones et [ksoten] ou [gzoten] par des francophones semble moins logique que *Khoten*. Parmi les Ouïghours même, certains maintiennent que l'utilisation de la lettre *x* complique les choses pour les habitants de la RAOX, qui doivent apprendre sa prononciation *pinyin* et une deuxième prononciation dans le cadre du LSU. Par contre, pour les ouïghours de l'ex-Union soviétique, cet usage rejoint celui du CSU et ne présente donc aucune difficulté.

ج [ʒ]: Parmi les 11 propositions pour ce son, aucune n'a totalement satisfait les exigences de tous les critères. Ici aussi, il y a eu un débat dû au fait que ce son n'existe dans les mots d'origine purement turque que comme une variante parlée de ج [dʒ] en position préconsonnantique (ex : [taʒliq] — qui porte une couronne, venant de [taɖʒ] — couronne),

⁸² La discussion a parfois pris un tournant amusant. Un exemple est la discussion autour de l'ambiguïté d'un mot épilé *haya* lorsqu'on utilise la lettre *h* indistinctement : ce mot pourrait donc transcrire et هايا [haja]—timidité et خايا [xaja], un mot plutôt cru pour désigner le pénis.

et est donc couvert par la lettre *j*. L'utilisation indépendante de ce phonème est limitée à quelques mots d'origine russe ou perse. Certains ont même estimé que l'utilisation de la seule lettre *j* pourrait rendre le LSU plus proche visuellement de l'usage international, puisqu'on la trouve dans des mots communs à beaucoup de langues (ex. *jandarma* — *gendarme*, français ; *jurnal*— *journal*, français et anglais, *injénér*— *engineer*, ingénieur, anglais ; *woltaj*— *voltage*). Mais ceci allait à l'encontre du principe A. De plus, cela privait le LSU d'un instrument pour transcrire des formes de langue non-standard dans la RAOX, tel l'ouïghour parlé dans URSS ainsi que les dialectes de Ghulja et Qeshqer. Il a donc été décidé de laisser entrer une double lettre, partiellement satisfaisante, **zh** choisie pour son utilisation dans les transcriptions occidentales de mots russes. Familière aux anglophones, cette forme est plus difficile à lire pour ceux qui parlent d'autres langues dans le monde. En plus d'être une lettre double, *zh* peut aussi prêter à confusion pour les Ouïghours de la RAOX puisque c'est aussi la forme *pinyin* utilisé en parallèle avec *j* dans le PNSU pour les mots d'origine chinoise (ex : *fangzhen* – directive). Les partisans de *j* n'ont pas été capables de convaincre ceux de *zh* et, puisque ce phonème ne se produit que rarement dans un contexte non dérivé, il a été décidé de l'utiliser à l'essai et de laisser l'usage décider.

ئا	ئە	ب	پ	ت	ج	چ	خ	د	ر	ز	ژ	س	ش	غ	ف
a	e	B	p	t	j	ch	x	d	r	z	j/zh	s	sh	gh	f
ق	ك	گ	ڭ	ل	م	ن	ھ	ئو	ئۇ	ئۆ	ۋ	ۋى	ئى	ئى	ي
q	k	G	ng	l	m	n	h	o	u	ö	ü	w	é	i	y

Table 3. Correspondance des lettres ASU et LSU.

La table ci-dessus n'inclut pas le hamza au dessus de la « dent » (ﺀ ou ﺀ), qui n'a pas été discuté au cours des conférences. Cette lettre n'est jamais listée séparément dans l'ASU, dû au fait qu'elle est considérée comme faisant partie intégrale de la forme initiale des voyelles⁸³. Mais, en fait, elle sert également à marquer en cours ou fin de mot un arrêt glottal. L'arrêt glottal n'est pas prononcé aussi fortement qu'en arabe ou en ouzbek, par

⁸³ Il se dit que la décision des linguistes ouïghours d'ajouter ce symbole comme faisant partie de la forme initiale des lettres est un lien à l'ancien système d'écriture ouïghour, dans lequel toutes les voyelles initiales étaient précédées d'une dent. L'alphabet arabe a trois lettres, و, ا, و qui peuvent indiquer les voyelles longues. Les voyelles courtes peuvent être indiquées par l'utilisation de marques au dessus ou en dessous des consonnes mais qui ne s'utilisent pas dans l'écriture courante. Etant donné ses caractéristiques phonétiques, le ouïghour prend en compte toutes les voyelles, ئى, ئى, ئى, ئى, ئى, ئى, ئى, ئى, ئى, ئى, ئى, ئى, ئى, ئى, ئى, ئى en utilisant des dérivés des lettres arabes.

exemple, et a été affaibli au point d'être rien de plus qu'une pause. En général il apparaît dans des mots d'origine arabe et remplace un 'ain (ع) ou un hamza (ء) dans la position médiane ou finale (ex. *سائت* de l'arabe *ساعة*, *خائن* de l'arabe *خائن*, *سوئال* de l'arabe *سؤال*, *ؤوقۇد* de l'arabe *وقوع*⁸⁴). Le « hamza au dessus d'une dent » utilisé dans un tel contexte, de manière traditionnelle, n'est pas vu comme étant une lettre différente mais comme étant une orthographe spéciale dans les positions médianes et finales.

Dans la plupart des transcriptions latines, la présence d'une pause (et donc d'un hamza par-dessus une dent dans un mot) est indiquée par l'usage d'une apostrophe. Le XSLC a choisi de suivre la même tradition pour ses règles d'orthographe. En même temps, il a été jugé non nécessaire d'en faire ainsi dans les cas où il y a deux voyelles séquentielles (ex. : *qaide*, *aile*, *saet*). L'apostrophe marquant une pause est donc réservé dans LSU aux cas où 1) une voyelle est séparée d'une consonne par une pause (ex. *qet'i*, *Qur'an*), 2) Les deux lettres *n* et *g* doivent être lues séparément et non comme [ŋ] dans *In'gliz*— anglais⁸⁵ ; (*baslan'ghuch*, *Hin'gan*), 3) Les deux lettres *s* et *h* doivent être lues séparément (*Is'haq*— Isaac); 4) Les deux lettres *ng* et *h* doivent être lues séparément (*Chong'haji*) et 5) la pause représente une ع ou ء finale dans des mots littéraires d'origine arabe (ex: *toghra'*— sceau royal, de l'arabe *طغراء*, *wuqu'*— phénomène, de l'arabe *وقوع*). Cet usage de l'apostrophe, sauf dans le dernier cas, ne respecte pas les règles actuelles d'orthographe de l'ASU mais au moins ne cause pas d'ambiguïtés.

2.6 Règles d'orthographe du LSU

1. La première lettre d'une phrase et d'un nom propre s'écrit en majuscule. Exemple : *Alim*, *Kanada*, *Uyghur*, *Ürümchi*.
2. L'orthographe des noms propres suit les règles de l'orthographe ouïghoure (comme si transcrit directement de ASU) mais l'orthographe originale peut être mis entre parenthèses

⁸⁴ L'arrêt glottal en position finale n'est généralement pas prononcé dans la langue moderne et n'apparaît que dans des textes classiques ou religieux.

⁸⁵ L'existence d'une seule lettre pour [ŋ] en ASU ئ, différencie ce son clairement du son *n+g* écrit avec deux lettres نى.

si nécessaire. Exemple : *Nyuyork* (New York), *Shenshi* (陝西/shǎn xī), *Skot Wéyd* (Scott Wade).

3. Les abréviations sont en majuscules, et les lettres séparés peuvent être suivies d'un point. Exemple : B.D.T (*Birleshken Döletler Teshkilati*, les Nations unies), TDMP(*Teywen Démokratik Musteqqilliq Partiyisi*, le Parti de l'Indépendance Démocratique de Taiwan).
4. Seule le premier élément d'une lettre composée est en majuscule. Exemple : AQSh (*Amérika Qoshma Shtatliri*, les Etats-Unis d'Amérique), Sh.U.A.R (*Shinjang Uyghur Aptonom Rayoni*).
5. Les unités de mesure et autres abréviations internationales ou marques déposées sont écrites telles que dans la langue d'origine, même si cela implique l'utilisation de lettres qui ne sont pas dans la liste des lettres du LSU. Exemple : *cm*, *kg*, *kv*, *VCD*, *MTV*, *Microsoft*, *Linux*.
6. Comme discuté ci-dessus, l'apostrophe indique une division en syllabes lorsque celle-ci n'est pas apparente, c'est-à-dire lorsqu'une voyelle ne forme pas une syllabe avec la consonne précédente. L'apostrophe est aussi utilisé afin de différencier la lettre *ng* du groupe de consonne *n + g* et la lettre *sh* du groupe de consonne *s + h*. Finalement, on utilise aussi l'apostrophe afin d'indiquer une présence étymologique d'un arrêt glottal final dans quelques mots classiques. A la différence du hamza en ASU, il n'est pas utilisé entre les voyelles afin d'indiquer une division entre les syllabes, puisqu'il ne peut y avoir qu'une voyelle dans une syllabe. Il faut noter que dans des mots directement importés du chinois et comportant les groupements phonétiques *-üen* (comme dans la devise chinoise *Yüen* يۈەن ou *shöyüen* شۆيۈەن , institut, et and *gowuyüen* گوۋۇيۈەن , conseil d'état) et *-hua* (comme *Jungxua* جۇڭخۇا , chinois) le ASU n'a pas de hamza intervocalique, transformant ainsi *ü* et *u* en semi-consonnes. Le traitement LSU de ces mots rétablit la prononciation de fait des Ouïghours, pour lesquels les semi-consonnes initiales *ü* et *u* n'existent pas⁸⁶.

⁸⁶ Les Ouïghours non sinophones ont tendance à prononcer ces mots *yen*, *shöyen*, *gowuyen* et *jungxuwa*

7. Des signes diacritiques doivent être utilisés lorsque requis puisqu'ils aident à rendre les sons plus exacts et à éviter les malentendus (*olturush* — s'asseoir; réunion, fête ; *öltürüsh*— tuer; tuerie, meurtre).

2.7 Statut actuel et perspectives futures

La version finale de l'alphabet LSU (voir table 3) a été rendue publique le 3 juillet 2001. Même si le LSU a été développé initialement pour un usage exclusivement informatique, il a rapidement été accepté par les Ouïghours éduqués et les ouïghourologues pour l'internet, ainsi que comme système de transcription dans les manuels et autres publications, où il a plus ou moins remplacé tout autre système de transcription.

Plus de sept ans d'expérimentation montrent une grande acceptation auprès des Ouïghours de la RAOX, les communautés ouïghoures à l'étranger et la plupart des ouïghourologues. Certains sites web ouïghours⁸⁷, afin d'attirer le plus de lecteurs possible, utilisent les trois transcriptions : ASU, CSU et LSU. Les outils⁸⁸ de conversion des écritures ouïghoures (ASU, CSU, LSU) se sont rapidement développés afin de permettre le partage d'information. Suite à cette évolution, les internautes ouïghours se sont mis à utiliser le LSU non seulement dans les courriels et les forums mais aussi dans le développement web, la recherche scientifique et dans le développement de logiciels d'applications. Les gens ont aussi commencé à réaliser que l'unification graphique fournie par LSU ne constitue pas une nouvelle réforme d'écriture imposée mais un instrument qui aide les Ouïghours à communiquer et à partager de l'information entre eux (Ouïghours de la RAOX, Ouïghours de l'ex URSS et Ouïghours installés dans d'autres pays), ainsi qu'avec d'autres locuteurs de langues turques apparentées à travers le monde.

Outre quelques incohérences de départ dues à des habitudes ancrées, particulièrement dans le passage du *x* du PNSU au *sh* du LSU, un problème majeur dans

⁸⁷ Voir biliwal.com, izdiyari.com, google.com/intl/ug/, rfa.org, oyghan.com, ukij.org, uyghurdictionary.org, <http://www.xjtsnews.com/normal/content/lading/index.htm> (site officiel du Bureau de la Propagande du gouvernement de la RAOX)

⁸⁸ Version démo disponible en ligne <http://www.uyghurdictionary.org/tools.asp>, version hors ligne ajoutciel pour Microsoft Word <http://oyghan.com/OTB/index.html>. Voir aussi www.ukij.org pour d'autres outils.

l'utilisation de LSU a été la réticence de certains utilisateurs d'ordinateurs à utiliser des diacritiques sur les voyelles *é*, *ö* et *ü*. Ceci vient surtout d'une « paresse du clavier » puisque, dans la plupart des cas, faute de claviers spécifiquement conçus pour le LSU, il faut une double manœuvre afin d'entrer les lettres accentuées (ex : avec le clavier international US : tréma/accent + lettre). Ce problème est en cours de résolution grâce au bon exemple donné dans les grands sites internationaux utilisant le LSU.

Un autre problème est que l'utilisation n'a pas totalement résolu le dilemme de *j* par rapport à *zh* dans le cas de la lettre ASU *Ĵ*, même si la plupart des ouïghourologues ont déjà opté pour l'usage de *zh*.

Chapitre 3. Traitement de la langue ouïghoure

Ce chapitre se concentre sur le traitement de texte et l’accessibilité de l’ouïghour sur le web. Il décrit le concept de base du développement d’une police ouïghoure basée sur l’Unicode, la méthode d’affichage et de saisie dans un environnement ne supportant pas l’ouïghour. On y souligne aussi les problèmes causés par l’absence de deux caractères ouïghours dans l’Unicode Standard, ainsi que les solutions proposées afin de surmonter ces lacunes.

3.1 Introduction

Officiellement, les Ouïghours de la RAOX utilisent l’alphabet arabe modifié⁸⁹ (ci-après : ASU) pour écrire leur langue, alors que l’ouïghour écrit en alphabet cyrillique modifié (ci-après: CSU) est toujours utilisé par les ouïghours de l’ex-Union soviétique. La translittération⁹⁰ récemment introduite – l’ouïghour en alphabet latin⁹¹ – a été reconnue et acceptée par les ouïghours et ouïghourologues. Cet alphabet latin est maintenant devenu un standard de translittération pour l’ASU et le CSU. Modifié afin de prendre en compte les caractéristiques phonétiques et grammaticales de la langue ouïghoure, l’alphabet arabe a été réintroduit chez les ouïghours de la RAOX en 1983, après une interruption⁹² de vingt ans. Dans ce chapitre, nous attirons l’attention sur quelques questions essentielles reliées au traitement de texte basé sur l’Unicode dans Windows en utilisant l’ouïghour écrit en alphabet arabe.

3.2 Historique

Au début des années 90, les développeurs de logiciels ouïghours ont mis au point une manière de faire des entrées de lettres ASU sous DOS. Puis à la fin des années 90, Microsoft Windows s’est répandu dans la société ouïghoure et beaucoup de logiciels complémentaires (logiciels *plug-in*, ou « ajoutciels ») ont fait leur apparition sur le marché afin de répondre aux besoins du marché local. L’absence de méthode d’entrée pour l’ouïghour et de police supportant tous les glyphes de l’alphabet ASU dans les plateformes

⁸⁹ Voir annexe 2

⁹⁰ On appelle translittération l’utilisation d’un alphabet pour en représenter un autre.

⁹¹ Voir annexe 10

⁹² Voir chapitre 3

existantes causait d'immenses problèmes dans l'affichage du texte et des caractères ouïghours dans différentes applications. Les développeurs locaux n'ont pas reconnu l'importance d'un standard commun pour tous les logiciels. Face à l'attitude passive en vigueur parmi les chercheurs parrainés par l'Etat et les autorités gouvernementales vis-à-vis du développement de toute technologie ayant trait à la langue ouïghoure, beaucoup de chercheurs individuels se sont lancés dans la création de logiciels et de sites web en ouïghour utilisant le LSU ou l'ASU. Pour créer leurs sites web, les webmasters ont surtout utilisé l'ASU en format image. Lorsqu'il a fallu passer au format texte, chaque site web a dû fournir sa propre police⁹³: les propriétaires des sites web ont ainsi créé et nommé des polices que les utilisateurs devaient télécharger en fonction des différents sites sur lesquels ils se rendaient. Même si les glyphes essentiels s'y trouvaient déjà, les développeurs de logiciels et de polices ne suivaient pas le standard Unicode en raison de difficultés sur lesquelles nous reviendrons en détail dans les sections qui suivent. Aucun webmaster n'était prêt à accepter les noms de police ni les règles d'encodage des autres, ce qui allait à l'encontre du besoin de standard commun. La création des polices se faisait par le remplacement de caractères ASCII ou des caractères arabes situés dans la zone 0x600-0x6FF de l'Unicode par des caractères ouïghours, sans que ces remplacements ne suivent de logique particulière: étant donné que le nombre des lettres arabes dans le code de gamme 0x600-0x6FF est plus important que celui des lettres ASU, chacun faisait un choix différent lors de la substitution d'un caractère ASU spécifique à un caractère arabe.

Chaque informaticien essayait de répondre à sa façon aux problèmes importants posés par l'utilisation de polices non standardisées. Ainsi un certain nombre de polices transitionnelles sans lien avec le standard Unicode ont été proposées pour contourner la difficulté. Comme le nombre potentiels des lettres arabes⁹⁴ « remplaçables » (c'est-à-dire qui ne sont pas utilisées dans l'ASU) présentes dans des polices au moment de l'installation du système d'exploitation est supérieur à celui des lettres ouïghoures absentes de l'alphabet arabe traditionnel, chacun proposait un « accord de remplacement » distinct qui substituait la même lettre ouïghoure à une lettre arabe « remplaçable » différente contribuant ainsi au

⁹³ Ci-après le terme « police ouïghoure » fait référence à une police basée sur l'ASU.

⁹⁴ Voir <http://www.unicode.org/charts/PDF/U0600.pdf>, <http://www.unicode.org/charts/PDF/UFB50.pdf>, <http://www.unicode.org/charts/PDF/UFE70.pdf> ; Arabe U0600-U06FF, Formes de présentation arabe A & B.

désordre qui régnait au niveau de la comptabilité. D'après un sondage effectué par l'auteur en décembre 2003, il y avait quelque 400 logiciels de traitement ASU en utilisation avec pas moins de 300 codages⁹⁵ de police. Voici un exemple de remplacement possible de lettres arabes. Notons que les formes et les codes des lettres⁹⁶ ouïghoures qui n'apparaissent pas dans la table suivante sont les mêmes que celles des lettres arabes et qu'aucune substitution n'est nécessaire dans leur cas:

14 Lettres arabes remplacées	ة	ص	ح	آ	ض	ث	ه	ذ	أ	ى	ؤ	ئ	ع	ظ
14 Lettres ouïghoures remplaçantes	ە	پ	چ	ژ	گ	ڭ	ھ	ۇ	ۆ	ۈ	ۋ	ې	ى	ۇ

Table 4. Une méthode transitionnelle: lettres arabes remplacées et ses remplaçantes ouïghoures

Pourquoi avoir eu recours à cette méthode transitionnelle ? Plusieurs facteurs peuvent être invoqués : 1) méconnaissance de la norme internationale — Unicode ; 2) complexité des tâches de développement des polices basées sur l'Unicode ; 3) problèmes d'indentification et d'ambiguïté des lettres ouïghoures dans la table de l'Unicode (nous reviendront sur ce point dans la section 4.3) ; 4) recherche d'un expédient pratique en attendant une solution durable⁹⁷ ; 5) ambition de monopolisation du marché ou autres motivations personnelles inconnues. Il faut admettre que le développement d'une police ouïghoure par cette méthode présente aussi des **avantages** : a) la modification d'une police « prête à remplacer » ne nécessite pas autant de travail qu'une création *Ex nihilo* ; b) Windows 95/98 vient avec une version antérieure à 1.47 du processeur multilingue — *the Uniscribe Unicode Script Processor* (USP10.DLL) qui est conçue pour le traitement de l'écriture arabe. Les règles de substitution définies dans ce processeur ne sont pas directement applicables à l'ouïghour et de ce fait ne sont pas compatible avec l'ouïghour. Mais, le remplacement de certaines lettres arabes par des lettres ouïghoures permet d'utiliser sans besoin de re-crédation les fonctionnalités de substitution des lettres arabes (qui

⁹⁵ Voir "Une proposition pour l'Unification des polices ouïghoures", <http://www.ukij.org/fonts/history/Teklip.pdf>.

⁹⁶ C'est dire: و، ن، م، ل، ك، ق، ف، غ، ش، ز، س، د، ر، خ، ت، ج، چ، خ، د، ر، ز، س، ش، غ، ف، ق، ك، ل، م، ن، و.

⁹⁷ Les réponses de quelques-uns de ces développeurs des polices non-standardisées (ou la méthode transitionnelle) que nous avons contactés étaient similaires : « c'est une solution temporaire. Comme dit le dicton, *At tapquche éshek min – avant de trouver un cheval, chevauche un âne.* ».

gèrent l’affichage des formes différentes des lettres) pour l’ouïghour selon les spécificités de l’écriture ouïghoure. L’inconvénient de ce système est que la création d’une méthode d’entrée par contrôle des événements (*keypress*, *keyup*, *keydown*) du clavier reste obligatoire pour effectuer des traitements (afficher, éditer etc.) de texte ouïghour, comme c’est le cas pour l’arabe ou d’autres langues à graphie non latine sur les anciennes versions des systèmes d’exploitation. Les **désavantages** considérables de cette méthode sont: a) non-conformité avec la norme internationale – Unicode, et de ce fait ; b) impossibilité de coexistence de l’ouïghour et de l’arabe dans le même texte en cas d’utilisation de la même police puisqu’elle implique le remplacement de plus de 40 glyphes de 14 lettres arabes; c) en l’absence d’accord de remplacement, on ne peut pas dissuader les développeurs de remplacer les lettres arabes de façon aléatoire.

La multiplication des noms des polices⁹⁸, les polices non standardisées basées sur la solution de contournement décrite ci-dessus et les méthodes d’entrées non-compatibles entre elles semaient la confusion parmi les utilisateurs, qui étaient en droit de se demander ce que faisaient leurs informaticiens. Plus encore, cette situation était nuisible à la diffusion de l’informatique et des instruments internet dans la population ouïghoure à un moment de grande progression de la sinisation de leur environnement. Finalement, il est devenu clair qu’il fallait impérativement mettre fin à cette situation extrêmement confuse en imposant un standard, Unicode ou non, pour le traitement de la langue ouïghoure.

En juin 2002, nous avons donc développé la première police ouïghoure basée sur l’Unicode et implémenté une méthode d’entrée ouïghoure pour Windows fonctionnant autant pour les systèmes d’opération que pour les navigateurs. Cette étape a été considérée comme révolutionnaire puisque cette nouvelle méthode permettait aux applications d’être totalement conformes au langage Unicode (contrairement aux anciennes méthodes qui ne l’étaient pas ou qui l’étaient seulement partiellement). Une campagne de lancement et de diffusion de cette police a ensuite été menée.

⁹⁸ Les mots clés les plus utilisés au début des noms de police sont : Uyghur XXX, Uighur XXX, le prénom des développeurs (ex. : Memet Asliya, Semet Tuz ...) et le nom des entreprises (ex. : Almas XXX, Alkatip XXX, UKK XXX)

En 2004, sous la direction des fondateurs de l'Association Informatique Ouïghoure⁹⁹, quatre conférences¹⁰⁰ ont eu lieu sur le net pour introduire le codage unifié de la police ouïghoure. L'auteur de cette thèse était chargé de l'organisation des événements et de la préparation des propositions¹⁰¹ initiales étudiées/discutées dans ces conférences. Parmi les participants, il y avait des experts en informatique travaillant pour des compagnies privées ; des organisations de recherche parrainées par le gouvernement ainsi que des développeurs de logiciels indépendants, à savoir Alim Ehed, Erkin Batur, Gheyret Toxti Kenji, Muhemmed Abdulla, Yasin Imin. Le but des conférences était de parvenir à une solution et à un commun accord sur les problèmes causés par certains glyphes ambigus de l'ouïghour. Dans les sections suivantes nous décrivons et soulignons les étapes essentielles à suivre lors de l'utilisation de l'ouïghour en Unicode pour le traitement de texte et le développement de logiciels.

3.3 Les lettres ouïghoures : problèmes d'identification et de caractères

L'ASU est une écriture complexe¹⁰² contenant 8 voyelles¹⁰³ et 24 consonnes (voir annexe 10), soit empruntées ou adaptées de l'alphabet arabe. Comme l'arabe, il s'écrit de droite à gauche et chaque lettre peut se former de deux à huit façons différents selon sa position dans un mot : les lettres ASU ont des formes initiales, médianes, finales et isolées ; certaines ont des formes conjointes¹⁰⁴. Au total, l'alphabet ASU a 126 glyphes différents.

⁹⁹ Connue en ouïghoure sous le nom *UKIJ – Uyghur Kompyutér Ilimi Jem’iyiti* en Uyghur), est une association but non lucrative, fondé par l’auteur de cette thèse en janvier 2004. Site web: <http://www.ukij.org>.

¹⁰⁰ Voir “l’histoire de l’unification des polices ouïghoures” <http://www.ukij.org/fonts>.

¹⁰¹ Voir <http://www.ukij.org/fonts> pour plus de détails.

¹⁰² Fait référence aux systèmes d'écriture nécessitant un degré de réorganisation et/ou traitement de glyphes afin de s'afficher, s'imprimer ou s'éditer. Voir www.microsoft.com/typography/Glyph%20Processing/intro.mspx pour plus de détails.

¹⁰³ L'alphabet arabe utilise 3 lettres pour indiquer les voyelles longues ou, dans le premier cas, utilisé comme soutien de voyelle و, ي, ا. Les voyelles courtes ne sont pas prises en compte dans l'écriture normale. Etant donné ses caractéristiques phonétiques, le ouïghour prend en compte toutes les voyelles : ئى، ئې، ئۆ، ئۇ، ئو، ئه، ئا, utilisant des dérivés de lettres arabes traditionnelles.

¹⁰⁴ La forme initiale et, sous certaines circonstances, la forme médiane de toutes les voyelles sont précédées d'un « panneau d'arrêt guttural » ʕ ou ʔ (hamza soutenu) avec lequel elles forment une lettre commune (traitée en ouïghour comme étant une lettre commune, voir annexe 10)

Comme en arabe, la combinaison de «ا» après «ل» nécessite deux glyphes de plus, selon sa position «لا» et «لا».

Les 108 glyphes de base¹⁰⁵ des lettres ASU ont été acceptées par l'ISO/Consortium Unicode. En 1998, 18 glyphes¹⁰⁶ ont été rajoutés en plus pour les formes conjointes des voyelles (voir ci-dessous, table 6). Puisque les 20 glyphes composés peuvent aussi être exprimés sous forme de deux glyphes déjà existants dans l'Unicode, les participants de la conférence mentionnée ci-haut ont suggéré de ne pas ajouter les 18 glyphes marqués en gris ci-dessous, dans la gamme de codage FBEA- FBEB. L'utilisation de ces 18 glyphes peut causer des problèmes tels la réduction de la rapidité de l'édition de texte, l'augmentation de la redondance de données stockées et compliquer les opérations de tri et recherche de données.

isolée'	finale'	médian'e	initiale'	isolée	finale	médiane	initiale
				ا	ا	ئا	ئا
				ه	ه	هه	هه
				و	و	مو	ئو
				و	و	مو	ئو
				و	و	مو	ئو
				و	و	مو	ئو
				و	و	مو	ئو
ي	ي	ي	ي	ئي	ئي	ئي	ئي
ي	ي	ي	ي	ئي	ئي	ئي	ئي

Table 5. Formes conjointes des voyelles ouïghoures (les positions primes, marquées d'une apostrophe, indiquent les positions où la voyelle n'est pas précédée de la marque d'arrêt glottal)

Malheureusement les formes conjointes médianes ¹⁰⁷ et ¹⁰⁸ des lettres ouïghoures ئي et ئي (indiquées en gris foncé ci-dessus) sont toujours absentes¹⁰⁹ de la table de

¹⁰⁵ Glyphes ne comprenant pas glyphes supportant le hamza. Voir annexe 7 et note 20.

¹⁰⁶ Voir *Arabic Presentation Forms-A*, zone : FBEA – FBFB. Voir aussi table **Erreur ! Document principal seulement.**, et unicode.org/charts/PDF/UFB50.pdf.

¹⁰⁷ Nom de caractère pour le standard Unicode: ARABIC LIGATURE YEH WITH HAMZA ABOVE WITH E MEDIAN FORM. Ex: باغېرىق (*Baghériq*).

¹⁰⁸ Nom de caractère pour l'Unicode: ARABIC LIGATURE UIGHUR KIRGHIZ YEH WITH HAMZA ABOVE WITH ALEF MAKSURA MEDIAN FORM. Ex: قەتئى (certainement)

¹⁰⁹ Les membres de la délégation de la ROAX, Prof. Hoshur Islam et Yasin Imin, qui ont envoyé la proposition admettent aussi cette omission. Voir aussi *Arabic Presentation Forms-A* (zone code: FBEA – FBFB).

l'Unicode Standard¹¹⁰ – *Arabic Presentation form – A*. Cette omission rend la présente table Unicode standard incomplète et impose l'« emprunt » de deux positions vides, FBD1 et FBD2. Afin d'éviter toute confusion, l'UCSA, qui compte soumettre une requête écrite au Consortium à ce sujet, a recommandé de supplanter ces deux glyphes dans toute version future de l'Unicode.

3.4 Autres sources de problèmes

Microsoft Office vient avec une police internationale nommée « Arial Unicode MS ». Elle inclut la quasi-totalité des formes (c'est-à-dire à part les deux qui sont indiquées ci-dessus) des lettres ouïghoures, mais ne contient pas la substitution de glyphes ni les caractéristiques de position des caractères. Nous pouvons dire la même chose pour certaines autres polices, telles « Times New Roman » et « Traditional Arabic », qui utilisent des séquences de substitution qui ont un affichage incorrect. Ex :

1. ئالەمدىكى ھەممە ئىنسان قەبىلى ھەممەس

2. ئالەمدىكى ھەممە ئىنسان قەبىلى ھەممەس

3. ئالەمدىكى ھەممە ئىنسان قەبىلى ھەممەس

(Tous les êtres humains du monde ne sont pas maléfiques)

Les phrases un et deux contiennent des combinaisons de caractères illégales s'ils utilisent les polices « Arial Unicode MS » et « Times New Roman », car les formes utilisées pour ئ, ھ, ق ne sont pas correctes d'après les règles de formation de lettres de l'alphabet ASU (voir annexe 10). Seule la phrase numéro 3 est correcte car elle utilise une police créée spécialement pour ASU, «UKIJ Tuz Tom ». Afin de créer les formes jointes correctes pour l'ouïghour, il est nécessaire de prendre des mesures spéciaux de traitement de glyphes pour les lettres à problèmes ئ, ھ, ق ainsi que les deux « marques d'arrêt » glottal, ٴ, ٴٴ. Ces problèmes peuvent être résolus en ajoutant toute l'information qui contrôle la substitution et le positionnement des glyphes ouïghours durant le développement des polices, notamment en utilisant les fonctions d'*Open Type Layout* de la technologie Open Type.

¹¹⁰ Voir <http://www.unicode.org/charts/PDF/UFB50.pdf>

Comme nous pouvons le constater, les lettres ouïghoures n'apparaissent dans l'ordre alphabétique ASU ni dans la table de code de caractères arabes ni dans la liste de noms de caractères du standard Unicode¹¹¹. Cela nécessite des traitements spéciaux lors de l'opération de tri (par ordre alphabétique). Les positions des lettres **ه**, **هـ**, **ي** sont ambiguës et les deux panneaux d'arrêt gutturaux **هـ**, **هـ** représentent la forme initiale et médiane de la lettre arabe **ي**, qui n'existe pas sous ses formes isolée et finale dans l'alphabet ASU. Nous les présentons dans la table ci-dessous avec le détail de leur position :

Formes représentatives	ي				ه	هـ				ئ	
codes formes	06CC ¹¹²		0649 ¹¹³		06D5 ¹¹⁴	0647 ¹¹⁵		06BE ¹¹⁶		0626 ¹¹⁷	
isolée	FBFC	ي	FEFF	ي		FEE9	هـ	FBAA	هـ	FE89	ئ
finale	FBFD	ي	FEF0	ي		FEEA	هـ	FBAB	هـ	FE8A	ئ
initiale	FBFE	هـ	FBE8	هـ		FEEB	هـ	FBAC	هـ	FE8B	هـ
médiane	FBFF	هـ	FBE9	هـ		FEEC	هـ	FBAD	هـ	FE8C	هـ

Table 6. caractères à problèmes dans la table de l'Unicode.

Dans l'Unicode, deux points de code différents sont donnés à chaque **ي** et **هـ**. La lettre arabe **هـ** *hah* a quatre formes différentes qui correspondent à quatre formes de deux lettres ASU différentes **هـ** et **هـ**. La lettre ASU **هـ** a un code unique, 06D5, mais les positions de ses quatre formes différentes ne sont pas indiquées dans les formulaires de présentation arabe. Dans l'alphabet ASU, les trois lettres **هـ**, **هـ**, **ي** ont les formes différentes suivantes (voir aussi annexe 10) :

¹¹¹ Voir <http://www.unicode.org/charts>.

¹¹² ARABIC LETTER FARSI YEH, formes initiales et médianes de cette lettre ont des points.

¹¹³ ARABIC LETTER ALEF MAKSURA, représente une lettre en forme de YEH sans points dans aucune position

¹¹⁴ ARABIC LETTER AE (Ouïghour, Kazakh, Kirghiz)

¹¹⁵ ARABIC LETTER HEH

¹¹⁶ ARABIC LETTER HEH DOACHASHMEE (Urdu)

¹¹⁷ ARABIC LETTER YEH WITH HAMZA ABOVE

isolée'	finale'	médiane'	Initiale'	isolée	finale	médiane	initiale
				ه	ه	ه	ه
ى	ى	ى	ى	ئى	ئى	ئى	ئى
				ھ	ھ	ھ	ھ

Table 7. Formes différentes de ه, ئه, ئى dans l'alphabet ouïghour (les positions primes, marquées d'une apostrophe, indiquent les positions où la voyelle ئى n'est pas précédée de la marque d'arrêt glottal)

Etant donné que les technologies d'Open Type permettent aux développeurs d'intégrer aux polices de l'information sur la substitution des formes (sélection automatique des formes), les décisions suivantes ont été prises par les participants de la conférence mentionnée ci-dessus et respectées lors de l'implémentation des polices dans la RAOX depuis fin 2004.

ئى: Lettre ouïghoure *i* comme dans *ishik* (ئىشك, porte). Elle a huit formes différentes comme décrit dans la table 8 ci-dessus. Pour les formes initiale', médiane', finale' et isolée', nous utilisons les quatre formes correspondantes (voir table 8) de la lettre arabe 0649 ي. Les formes finale' et isolée' ne devraient pas être confondues avec les formes finales et isolées de la lettre Farsi TET 06CC. Les formes conjointes devraient être exprimées comme la combinaison des formes initiale et médiane de la lettre arabe YEY 0626 ي, suivies des formes médiane et finale de la lettre arabe 0649 ي. Par exemple: l'engin de formation des positions de police Open Type (*Open Type font shaping engine*: une technologie à l'intérieur de la police qui aide à sélectionner les formes correspondantes – initiale, médiane, finale et isolée) devrait combiner ي (FEF0) avec ي (FE8B) afin d'obtenir ئى.

ه: Lettre ouïghoure *e* comme dans *eyneklerde* (ئەينەكلەردە, dans les miroirs). Cette lettre utilise les formes finales et initiales (ه, ه) de la lettre arabe 0647 ه¹¹⁸, tout comme le perse. Cela cause un problème spécial puisque les glyphes du même caractère arabe ه dans les positions initiale et médiane (ه, ه) correspondent à ceux du ouïghour ه (*h* comme هېلەم *hélihem*, même maintenant; گۇناھ *gunah*, péché ou offense; قەبىھ *qebih*, odieux), qui a des

¹¹⁸ Voir aussi <http://www.unicode.org/standard/where/>, Formes variées de la lettre arabe *hah*.

formes finale et isolée différentes (ھ , ھ, voir table 8). Afin de contourner cette incohérence, nous avons choisi d'utiliser 06D5 pour la lettre ouïghour ھ et 06BE pour la lettre ouïghoure ھ. De plus, la lettre ourdoue 06BE a quasiment les mêmes formes que la lettre ouïghoure ھ. À l'intérieur des polices ouïghoures, les recherches de substitution de 06D5 utilisent FEE9 et FEEA dans les positions isolée et finale de ھ. Les formes composées ھ et ھ peuvent être obtenues en rajoutant la forme finale de ھ 06D5 aux formes initiale et médiane de 0626 ئ, comme nous le faisons pour ئى.

ئ et ئ¹¹⁹: « marques d'arrêt glottal ». Ceci est un phonème qui n'est pas marqué séparément dans l'alphabet ASU mais qui dépend quand même de son orthographe. Cet arrêt guttural n'est pas prononcé aussi fortement en ouïghour qu'en ouzbek ou dans les langues sémitiques par exemple et il a été affaibli au point de ne plus être qu'une pause. Marqué dans l'ASU par un hamza au dessus d'une « dent », il apparaît généralement dans des mots d'origine arabe où il remplace un 'ain (ع) ou un hamza (ء) en position médiane ou finale (ex. سُؤَال de l'arabe سُؤَال, خَائِن de l'arabe خَائِن, سَاعَة de l'arabe سَاعَة, عَالَم de l'arabe عَالَم). En position initiale, ce même symbole est considéré comme faisant partie de la forme initiale d'une voyelle et n'a aucune valeur phonétique¹²⁰. Dans l'Unicode, les marques d'arrêt glottal correspondent aux formes initiale et médiane de la lettre arabe 0626 ئ. Ces glyphes arabes ne sont pas considérés comme étant des formes différentes d'une quelconque lettre de l'alphabet ouïghour (cf. annexe 10). Puisque un glyphe de chacune des deux lettres ئى and ئى (en gris foncé dans la table 1) manque toujours dans l'Unicode, nous pouvons utiliser une séquence d'un des glyphes (ئ ou ئ) suivis des formes finale, isolée, médiane ou finale des voyelles ئى et ئى. Par conséquent, toutes autres formes composées des voyelles (cf. table 6) peuvent être obtenues en joignant une forme représentative d'une voyelle à la lettre arabe 0626 ئ.

¹¹⁹ Formes initiales et médianes de 0626 ئ.

¹²⁰ On dit que la décision des linguistes ouïghours d'ajouter ce symbole à la forme initiale des voyelles est un lien avec l'ancien système d'écriture ouïghour, dans lequel toutes les voyelles initiales étaient précédées d'une « dent ».

Malgré les limitations mentionnées ci-dessus (deux glyphes au lieu d'un glyphe composé pour ئى et ئې), les conventions mentionnées ci-haut ont été bien acceptées par l'Association Informatique Ouïghoure et par la filiale de l'Université du Xinjiang du Groupe de Recherche 863¹²¹. L'annexe 7 donne plus de détails sur les formes représentatives et les codes de glyphes pour la formation des lettres lors de la substitution et du positionnement.

3.5 Développement des polices ouïghoures et une méthode d'entrée au niveau du système d'exploitation

Après avoir maîtrisé les spécifications des lettres problématiques décrites ci-dessus, il est possible de créer des polices en utilisant les valeurs conventionnelles des lettres ouïghoures (voir annexe 7) avec des logiciels existants de création de police. L'inclusion des marqueurs sans-espaces, tels ZWJ (*zero width joiner* 200C), ZWNJ (*zero width non-joiner* ; 200D), LTR (marqueur de gauche à droite ; 200E) et RTL (marqueur de droite à gauche ; 200F), est recommandé dans les polices ouïghoures. Le reste de la tâche répétitive du développement de police ressemble tout à fait aux tâches de la création de polices arabes¹²². L'auteur a créé la première police ouïghoure basée sur l'Unicode en 2002 et l'a distribuée gratuitement sur l'internet afin de contribuer à la standardisation des échanges d'information en langue ouïghoure. De ce fait, quelques polices ouïghoures basées sur les conventions mentionnées ci-dessus sont disponibles gratuitement sur le site web de UCSA¹²³. Nos outils de création de police recommandés sont : *Font Creator*¹²⁴, *FontLab*¹²⁵ et *Fontographer*¹²⁶.

¹²¹ Un groupe de recherche national sur la haute technologie, financé par le gouvernement de la RPC. La filiale XJU est spécialisée dans le développement de logiciels multilingues.

¹²² Voir <http://www.microsoft.com/typography/OpenType%20Dev/arabic/intro.msp> pour plus d'informations sur le développement de polices Open Type pour l'alphabet arabe.

¹²³ Voir <http://www.ukij.org/fonts>

¹²⁴ Voir <http://www.high-logic.com/fontcreator.html>

¹²⁵ Voir <http://www.fontlab.com>

¹²⁶ <http://www.fontlab.com/Font-tools/Fontographer>

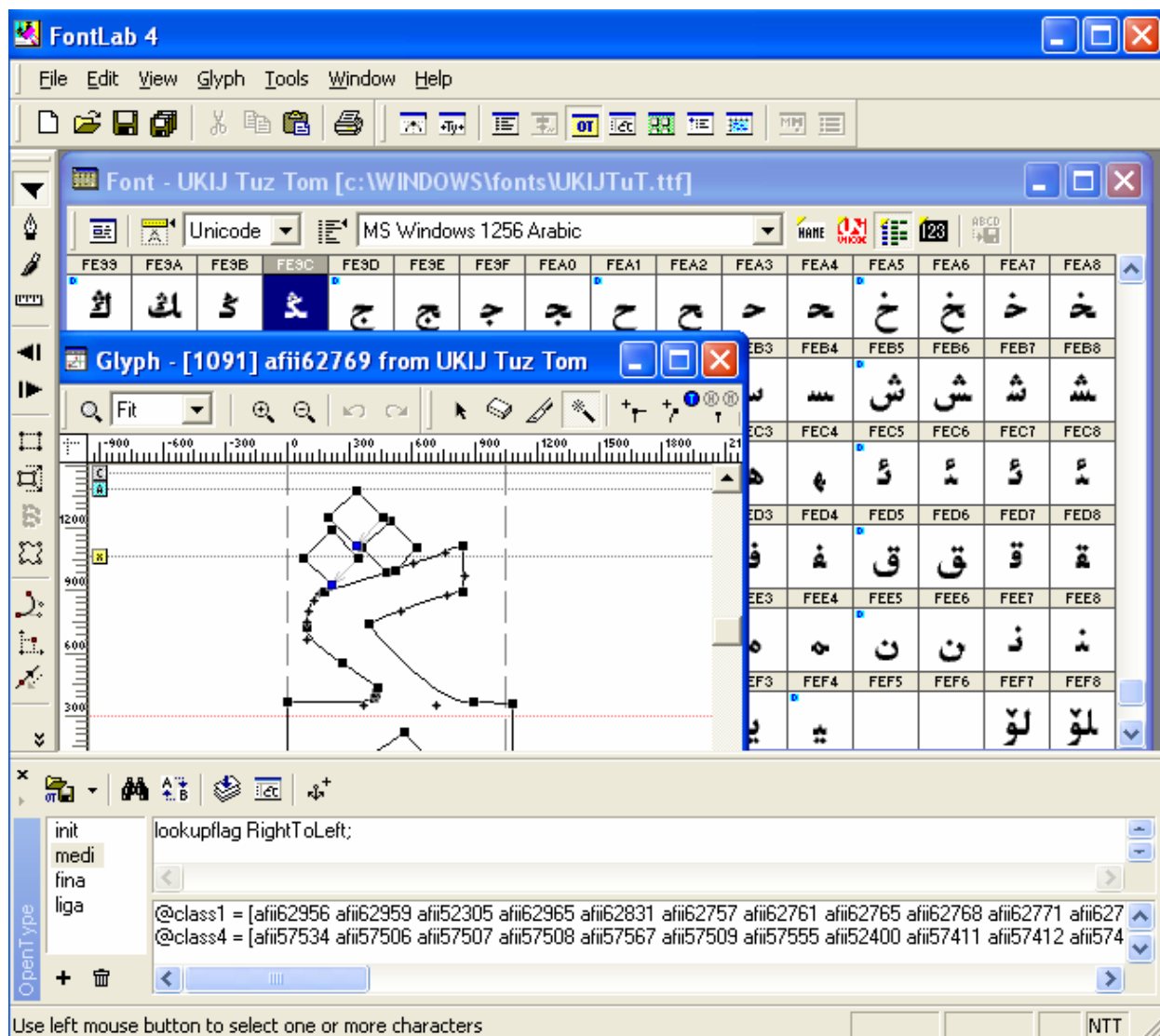


Figure 1. vue d'une lettre ouïghoure lors de la création de police sur l'outil – Font Lab.

L'intégration des fonctionnalités de substitution est une des étapes clés du développement d'une police. Si celle-ci est absente de la police, l'affichage des lettres se fait incorrectement. Par exemple :

1. ئاپتوماتىك شەكىل تاللاش مىسالى

2. ئاپتوماتىك شەكىل تاللاش مىسالى

Exemple de substitution des formes

La ligne (2) représente le même texte que la ligne (1) tel qu'il s'affiche en l'absence de fonctionnalité de substitution dans la police. On peut voir clairement que les lettres n'y

sont pas connectées, ce qui est incorrect. Les substitutions de glyphes, recherches de positionnement, fonctions de formation et les tables Open Type de polices arabes peuvent aussi être ajoutées grâce à des logiciels tel Microsoft VOLT et FontLab.

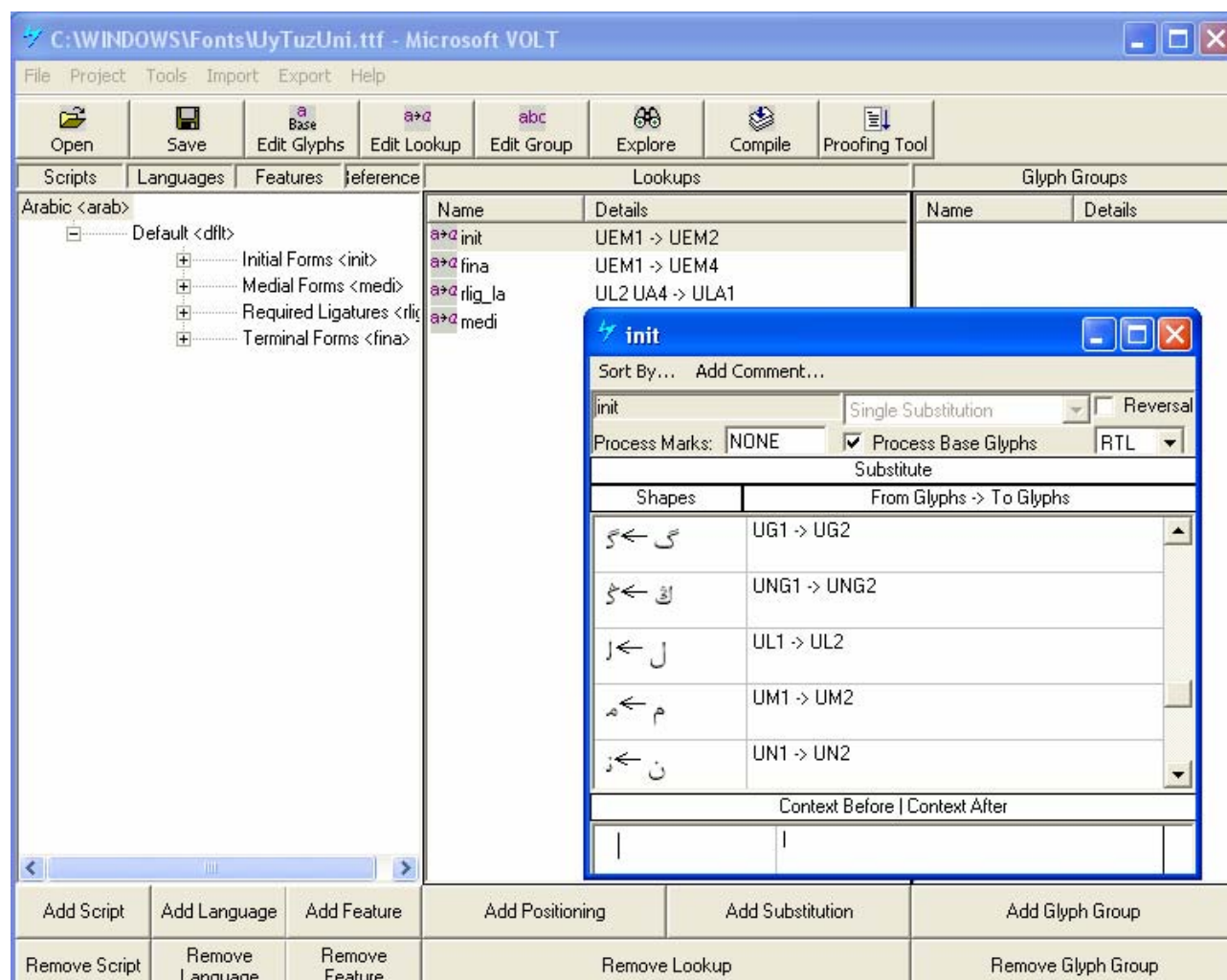


Figure 2. Exemple d'intégration des fonctionnalités de substitution

Supposons maintenant que nous avons déjà des polices ouïghoures développées en sur la base de l'Unicode et des conventions mentionnées ci-dessus. Puisque les systèmes d'exploitation existants ne fournissent pas¹²⁷ une méthode d'entrée (désormais : IME) pour l'ouïghour, nous avons encore besoin de créer des outils permettant de saisir les lettres ouïghoures. Il y a deux types de méthodes d'entrées:

¹²⁷ À part Windows Vista, dans lequel subsiste cependant un problème de compatibilité de la lettre « ف ». D'après Microsoft, ce problème sera corrigé dans la prochaine version de Windows (sous forme de patch.)

- A. **Méthode traditionnelle** : cette méthode nécessite la présence des instructions de substitution (voir figure 1 et 2) dans la police, et une version ultérieure à 1.47 du processeur multilingue Uniscribe (USP.dll) qui les interprète/décodes afin de sélectionner des formes correctes parmi les formes possible (initiale, médiane, finale, isolée) en suivant les règles de l'alphabet ouïghour. Lors de la saisie, la méthode d'entrée n'a besoin que de transmettre des codes situés dans la zone de base Unicode (Basic Unicode Range), soit des valeurs entre 0600 – 06FF. Si cette méthode est utilisée, le mot « ئاپتوماتىك » devient « ئ ا پ ت و م ا ت ى ك » lorsqu'on met un espace entre chaque lettre. Ce qui indique que la substitution s'est faite pendant l'affichage et que dans la mémoire il n'y a que les formes représentatives des lettres saisie. Voir table 10 pour les lettres représentatives et leurs codes.
- B. **Méthode étendue** : cette méthode envoie des codes situés dans la zone d'extension – *Arabic presentation forms A & B* (soit des valeurs : FE70–FEFF, et FB50–FDFF) de l'Unicode. Les lettres passent par un processus d'entrée qui sélectionne la bonne forme parmi les formes possibles pour chacune des lettres avant de l'afficher à l'écran. Si cette méthode est utilisée, le mot « ئاپتوماتىك » devient « ئ ا پ ت و م ا ت ى ك » lorsqu'on met un espace entre chaque lettre. Ce qui indique que la substitution se fait avant l'affichage et dans la mémoire il y a un code pour chacune des « formes phasiques ». L'utilisation des « formes phasiques », comparativement aux formes représentatives, complique la tâche de contrôle des formes dans la mémoire. Voir table 10 pour les formes physiques et leurs codes.

Méthode trad.	lettres	ك	ى	ت	ا	م	و	ت	پ	ا	ئ
	codes	0643	06CC	062A	0627	0645	0648	062A	067 ^E	0627	0626
Méthode étendue	formes	ك	ى	ت	ا	م	و	ت	پ	ا	ئ
	code	FEDA	FBE7	FE97	FE8E	FEE3	FEEE	FE98	FB58	FE8E	FE8B

Table 8. deux types de méthode d'entrée

Dans la méthode traditionnelle, les formes initiale et médiane de la lettre «ت» ont toujours le même code (U062A) dans le mot « ئاپتوماتىك ». Mais elles auront deux codes différents (FE97, FE98) en cas d'utilisation de la méthode étendue qui ne dépend pas du

processeur multilingue Uniscribe (USP.dll). Le seul avantage de la méthode étendue est qu'une fois les lettres saisies elles peuvent être affichés correctement même dans des logiciels qui ne supportent pas l'Unicode. C'est pour cela que la méthode étendue était populaire sous Windows 95/98. Cependant, les difficultés persistent lors de l'insertion et du tri en raison de la présence des formes physique.

Windows 2000 et les versions ultérieures contient un processeur multilingue — *the Uniscribe Unicode Script Processor* (USP10.DLL), qui permet de manipuler correctement les positionnements des formes différentes des lettres des écritures complexes¹²⁸ aussi bien au niveau de l'application qu'à celui du système. D'ailleurs, le soutien multilingue pour les langues écrites de droite à gauche, présent dans Windows 2000 et ses versions ultérieures, favorise le traitement de texte en ouïghour basé sur l'Unicode. Une combinaison de ces deux éléments peut être très utile dans le développement d'une méthode d'entrée traditionnelle qui fonctionne au niveau de l'application et du système. 1.471 ou des versions plus récentes de USP10.DLL avec un moteur de positionnement qui appuie sans réserve l'ouïghour. Fin 2003, nous avons développé la première méthode d'entrée basée sur l'Unicode et nous l'avons distribué gratuitement sur l'internet¹²⁹. Six mois plus tard, la filiale du Groupe de Recherche 863 à l'Université du Xinjiang et quelques chercheurs isolés se sont joints à la campagne de la popularisation de l'Unicode ouïghour en distribuant leurs IME commerciaux. Aujourd'hui, notre IME est devenue un des IME gratuits les plus utilisés dans la communauté internautes ouïghoure. Voici quelques copies d'écran qui peuvent aider à mieux voir comment s'est faite l'intégration au système:

¹²⁸ Voir note 106 écriture complexe.

¹²⁹ Disponible sur <http://www.ukij.org/oyghan/unicode/UyghurUnicodeIME.htm>

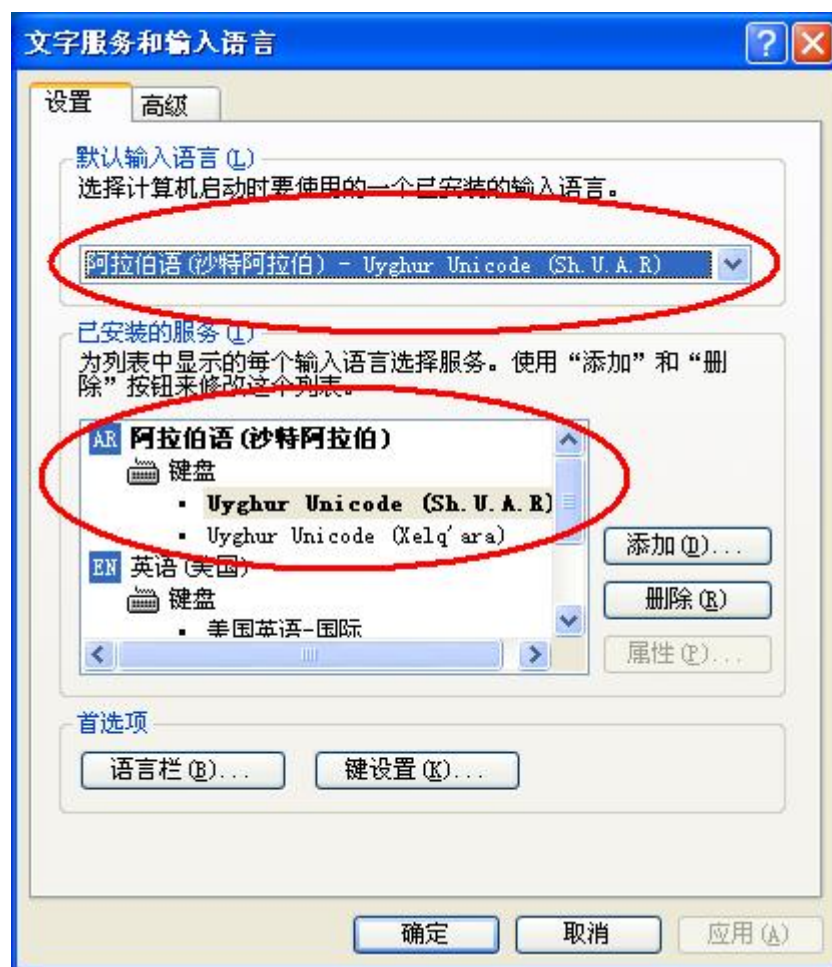


Figure 3. Vue de la fenêtre des services de texte et langues d'entrées après l'installation de notre IME¹³⁰.

Nous avons fourni deux dispositions du clavier ouïghour. La première – «Uyghur Unicode (Sh.U.A.R)» sur la figure 3 correspond à la norme régionale du RAOX, pour le clavier QWERTY.



Figure 4. disposition du clavier ouïghour (norme régionale)

¹³⁰ Extrait d'une image du manuel d'installation de notre IME. Plus d'images disponibles dans le manuel en ligne : <http://www.ukij.org/oyghan/unicode/UyghurUnicodeIME.htm>. Comme la grande majorité des ouïghours utilise la version chinoise de Windows, nous avons utilisé des copies d'écran chinois dans le manuel.

La disposition ci-dessus avait été conçue à la fin des années 80s à l'intention des professionnels de la capture texte plutôt que pour le grand public. Afin d'augmenter la vitesse de frappe, les lettres les plus fréquentes sont placés là où il y a le minimum de mouvement des doigts une fois les deux index sur les touches F et J et les autres doigts sur « D S A » et « K L ; ». Il y a deux lettres sur chacune des touches entre « D » et « L » incluses. Avec cette disposition, il faut frapper sur la touche « a » pour saisir « ھ » (*h* → *équivalent en LSU*), « f » pour saisir « ڭ » (*a*), « g » pour saisir « ە » (*e*) ...Au total 9 lettres dont la prononciation n'a rien avoir avec les lettres marquées sur les touches. Les autres touches avaient été sélectionnées par rapprochement, autant que possible, avec la prononciation en *Pinyin* des lettres figurant sur les touches. Naturellement, ceci n'est guère encourageant les analphabète de l'informatique. Aussi, nous avons proposé, pour la première fois, une autre disposition du clavier – appelée « Uyghur Unicode (Xelq'ara → Internationale) », dont la conception est inspirée de l'écriture latine-ouïghoure (LSU).

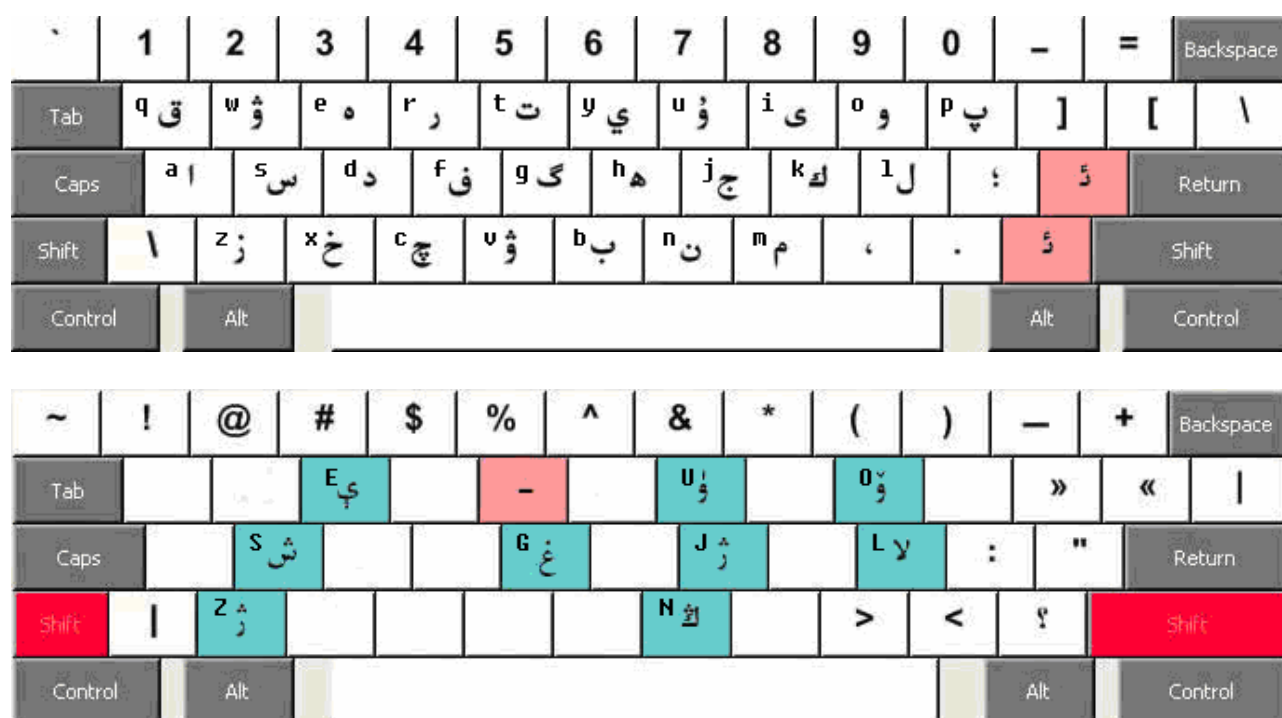


Figure 5. disposition du clavier ouïghour basée sur le LSU

Cette disposition du clavier est relativement facile à retenir du fait de la ressemblance phonétique entre les lettres latines sur les touches du clavier QWERTY et les

lettres ouïghoures en ASU. Dans cette disposition, il faut frapper sur la touche « a » pour saisir la « ا » ($a \rightarrow \text{équivalent en LSU}$), « f » pour saisir « ف » (f), « e » pour saisir « e » (e) etc... ; « ö, ü, é » sont placées sur les touches « O, U, E » (leurs équivalent en LSU sans les diacritiques) ; « sh, gh, ng » sont placées sur « S, G, N » (leurs équivalent en LSU en cas absence des deuxièmes lettres – h, h, g) ; deux possibilités pour « J, Z » comme c'était le cas dans l'alphabet LSU ; deux possibilités également pour « w, v » car « w » et « v » sont prononcés de la même façon en ouïghour, même si « v » n'est pas dans l'alphabet LSU. Cette disposition est préconisée par des « nouveaux apprentis », mais contestée par des gens qui sont contre l'utilisation de la LSU car « cela accélère l'apprentissage de l'alphabet LSU, ce qui représente un danger pour la survie de la ASU ».

Même en l'absence de l'ouïghour dans la définition des *locales*¹³¹ dans Windows XP et les versions plus anciennes, l'ajout d'un IME ouïghour ne se heurte à aucun obstacle tant qu'il est relié à un IME permettant d'écrire des lettres arabes. Nous avons choisi de relier notre IME au IME de l'Arabie Saoudite (voir figure 3), sans raison particulière, puisque nous aurions aussi bien pu le faire avec n'importe quel autre IME arabe sans que cela ait de s conséquences particulières. Une fois l'IME ouïghour installé, nous pouvons saisir des lettres ouïghoures dans n'importe quel endroit qui permette de saisir du texte en choisissant une des deux dispositions du clavier proposées.



Figure 6. IME ouïghour sur la barre de langues Windows

Nous présentons ci-dessous le procédé de contrôle qui capture et modifie les événements de clavier d'entrée les caractères ouïghours :

¹³¹ aussi appelés paramètres régionaux. « Ce sont un ensemble de définitions qui permettent au logiciel d'afficher les données selon les attentes culturelles et linguistiques propres à la langue et au pays de l'utilisateur, à savoir: le type d'écriture, le type de virgule, la représentation des chiffres, le format de la date et de l'heure, les unités monétaires, l'encodage par défaut, l'ordre alphabétique des lettres (qui peut différer selon les régions); ». voir : <http://fr.wikipedia.org/wiki/Locales>

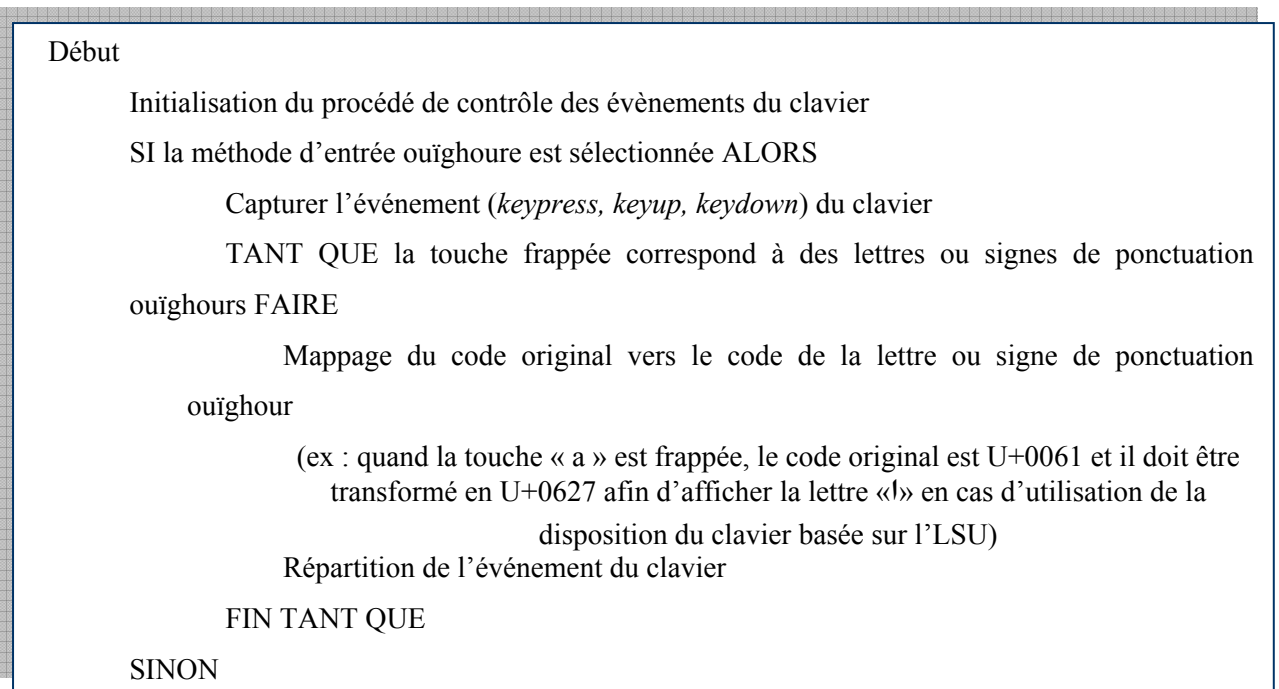


Figure 7. Algorithme de la méthode d'entrée ouïghoure au niveau du système

Plus de détails sur l'affichage des lettres, le développement d'un IME au niveau du navigateur et les outils de conversion multilingues sont présentés dans les sections suivantes.

3.6 Incorporation des polices et l'affichage des lettres ouïghoures

Les sites web peuvent être rendus sans télécharger ou installer de polices spécifiques si :

- 1) La police utilisée dans les pages est disponible dans l'ordinateur de l'utilisateur et
- 2) Si le navigateur fournit un soutien des langues natives et polices utilisées.

La deuxième condition a déjà été remplie mais malheureusement ce n'est pas le cas pour la première, puisque il n'existe aucune police ouïghoure disponible à travers les plateformes installées dans les ordinateurs des utilisateurs. Par conséquent, afin de veiller à ce que les polices ouïghoures soient affichées correctement dans les navigateurs Web, les utilisateurs doivent trouver une façon d'installer dans leurs ordinateurs les polices qui y sont utilisées. Ceci est le cas pour toutes les autres « langues oubliées » dans d'autres plateformes. L'exigence de l'installation de police cause des difficultés auprès de ceux qui n'ont pas beaucoup d'expérience technique et peut décourager d'autres de tenter de lire le texte.

Ces difficultés peuvent être surmontées par l'inclusion des polices dans les pages web. Quand une page est chargée dans un navigateur via le protocole de transfert hypertexte, les polices intégrées dans la page sont également téléchargées sans qu'il soit nécessaire pour l'utilisateur d'intervenir. Avec WEFT¹³², il est possible de transformer une police TrueType ou OpenType en police incorporée, qui peut par la suite être intégrée dans des pages web sous forme d'un objet. Les étapes suivantes permettent aux développeurs de créer des pages Web qui peuvent afficher des lettres ouïghoures correctement sans intervention de l'utilisateur pour installer des polices utilisées dans des pages:

- Préparer les pages web en utilisant n'importe quelle police installée sur la plateforme;
- Créer des polices incorporées pour chacune des polices utilisées dans les pages web en utilisant Microsoft WEFT ;

A la fin de la procédure de transformation des polices TrueType ou OpenType en polices incorporées, WEFT génère un fichier avec une extension (.EOT). Ce fichier doit être intégré sous forme d'un objet dans des pages web.

- Intégrer les polices incorporées dans les pages web.

WEFT génère également un code, sous la forme suivanet, qui permet d'intégrer les polices incorporées (le fichier avec l'extension .EOT) dans les pages web. Il suffit de le mettre entre les balises <body> ... </body> dans le code HTML.

```
<STYLE type=text/css>
    @font-face {
        font-family: UKIJ Tuz;
        src: url(UKIJTUZ0.eot);
    }
</STYLE>
```

L'inconvénient des polices incorporées générées par WEFT est que les polices sont compatibles uniquement avec Internet Explorer. Nous suggérons fortement d'investir plus d'efforts afin de fournir une compatibilité inter-plateforme pour ce genre de logiciel.

¹³² Logiciel gratuit de Microsoft, disponible sur : <http://www.microsoft.com/typography/web/embedding/default.htm>

3.7 Création d'une méthode d'entrée virtuelle au niveau des navigateurs

Comme nous l'avons mentionné dans l'introduction, les plateformes existantes ne fournissent aucune méthode d'entrée pour la langue ouïghoure au niveau du système. Bien que nous ayons développé un IME au niveau système, nous ne pouvons pas dire que tous les internautes ouïghours sont équipés de cet outil. Donc, la méthode d'entrée au niveau du navigateur est encore très en demande puisqu'elle permet d'insérer n'importe quel caractère en ouïghour dans les champs textes d'un site web sans avoir à installer un IME ouïghour au niveau du système. La structure de base de cet outil d'entrée au niveau de l'explorateur est représentée dans la figure ci-dessous:

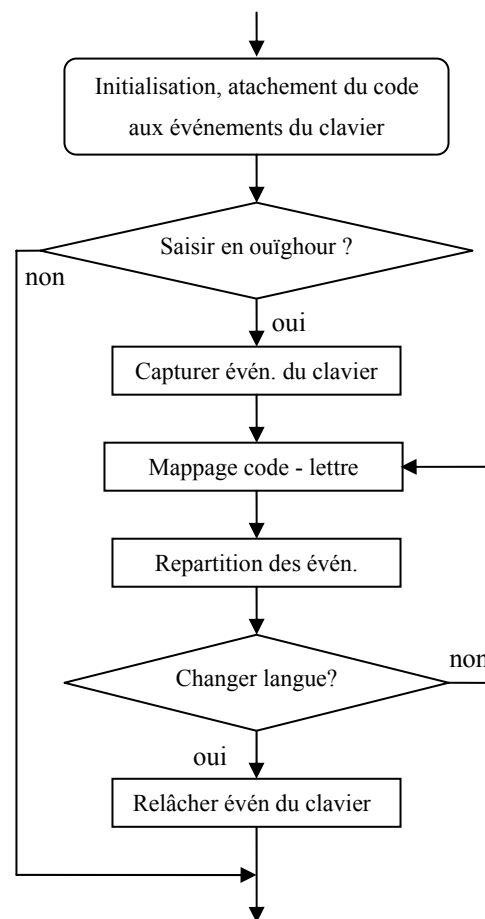


Figure 8. Schéma de travail de la méthode d'entrée au niveau des navigateurs

Comme nous pouvons le constater par le schéma ci-dessus, la structure générale de cette méthode ressemble à celui de la méthode d'entrée au niveau du système mentionnée dans la figure 7. Une fois que l'utilisateur choisit l'option entrée ouïghour, le module

« Capturer événements du clavier » crée un crochet afin de surveiller les activités du clavier telles que *keypress*, *keydown*, *keyup*. Le module « mappage code-lettre » crée une matrice de keycode-vers-unicode afin d'obtenir la lettre ouïghour correspondant au code de la touche (ex : U+006D → م) selon la disposition du clavier choisie. Le module « Répartition des événements » envoie les lettres ouïghours au champ d'entrée de texte sur une page web. Ce processus se répète jusqu'à ce que le module « relâcher événements du clavier » libère le crochet, immédiatement après que l'utilisateur a choisi de passer de la méthode d'entrée ouïghoure à une autre. Cette méthode avait été implémentée en utilisant VBScript en 2002 puis distribuée sur l'internet sous forme de logiciel Open Source. Après plusieurs modifications, une version Javascript a été mise en œuvre par d'autres développeurs afin de la rendre compatible avec d'autres navigateurs qu'Internet Explorer. Cette méthode est testée sur des navigateurs différents communément utilisés dans certains sites web ouïghours¹³³. Nous présentons son utilisation dans la section évaluation de ce chapitre.

3.8 Conversion multi-écriture

Puisque la langue ouïghoure utilise trois systèmes d'écriture (alphabets arabe — ASU, cyrillique — CSU et latin — LSU), nous avons fait des recherches visant à créer un outil de conversion permettant d'effectuer la conversion entre les trois écritures dans le but de faciliter les échanges d'information.

Le fait qu'il y ait une correspondance entre les lettres de ces trois systèmes d'écriture est certainement un facteur positif. Afin de mieux comprendre, prenons l'exemple du proverbe ouïghour, « travailler pour rien est mieux que de ne rien faire » dans les trois alphabets :

بىكار يۈرگىچە بىكار ئىشلە

бикар йүргичә бикар ишлә

bikar yürgiche bikar ishle

Le schéma suivant explique le processus de la conversion de base :

¹³³ Voir www.ukij.org, www.biliwal.com, www.oyghan.com, www.uyghurdictionary.org etc.

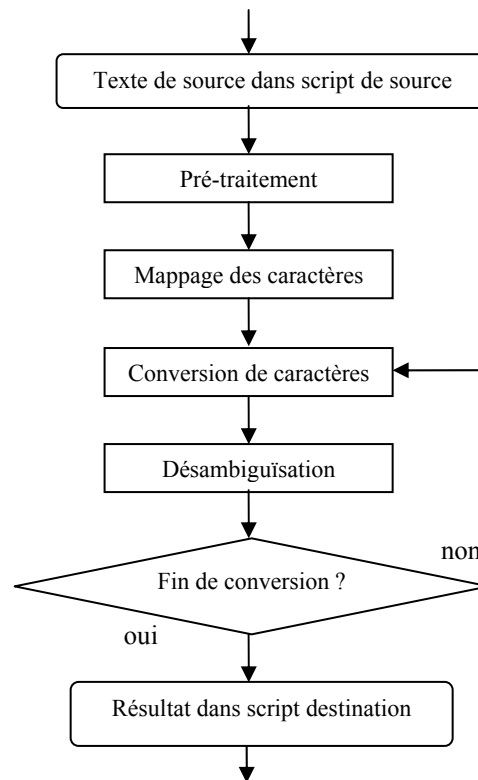


Figure 9. structure du convertisseur multi-écriture

Il se peut que les fonctionnalités de chaque module doivent être clarifiées :

Pre-traitement : une étape très importante dans la conversion. Elle inclut préserver des éléments qui devraient rester inchangés¹³⁴ après la conversion. Par exemple, lors que nous convertissons le texte LSU : “*Men Photoshopni yaxshi körimen*” (J’aime Photoshop) en ASU, nous devrions obtenir : “*من Photoshop نى ياخشى كۆرىمەن*” et vice versa. **Mappage des caractères**: crée une matrice du style “*B_devient_ب*” matrice pour chaque combinaison d’écritures Un total de trois matrices sera créé : LSU $\leftrightarrow$ ASU, CSU $\leftrightarrow$ ASU, CSU $\leftrightarrow$ LSU. **Conversion de caractères**: utilise les trois matrices afin de convertir entre les trois alphabets. **Désambiguïsation**: Ce module est nécessaire lors de la conversion de LSU à ASU et/ou CSU, à cause des fautes d’orthographe ou à cause de problèmes dus à la difficulté de taper les diacritiques du LSU sur certains claviers. Fréquemment, les lettres *Ö, Ü, É, ö, ü et é* sont remplacées par *O, U, E, o, u et e*. Ceci peut causer des erreurs fatales. Par exemple : *öltürüşh* (tuer) $\Leftrightarrow$ *olturush*(s’asseoir, faire la fête), *térím yer* (terre fertile) $\Leftrightarrow$ *terim yer* (qui mange ma transpiration), *yétim*(orphelin) $\Leftrightarrow$ *yetim*(faute d’orthographe).

¹³⁴ C’est le cas pour les liens hypertext, les tags HTML et les noms propres.

D'ailleurs, les fautes d'orthographe causés par le non respect des règles du LSU sont un problème assez fréquent. Il est donc souvent impératif de passer par une étape de correction d'orthographe. Cette fonctionnalité de l'outil¹³⁵ de conversion multilingue que nous avons publié sur internet est encore en cours de développement.

Nous avons implémenté deux outils de conversion basés sur la structure décrite dans la figure 9 : un barre d'outil pour Word¹³⁶ et un outil en ligne. Les images suivantes aideront à mieux comprendre les fonctionnalités de ces outils :

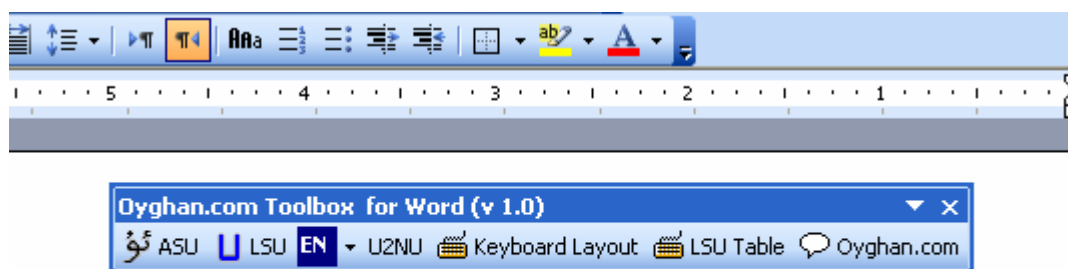


Figure 10. Vue du convertisseur multi-écriture sur Microsoft Word.

Pour convertir d'une écriture vers l'autre, il suffit de cliquer sur le bouton qui appelle le procédé de conversion :

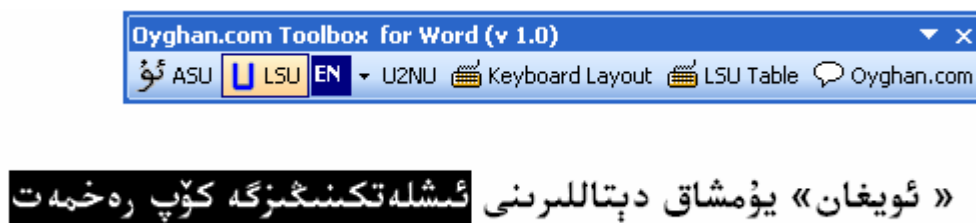
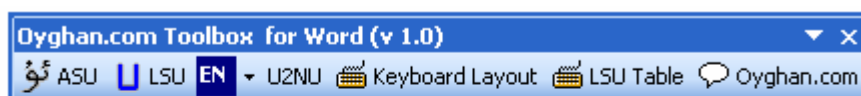


Figure 11. vue avant conversion du texte sélectionné de l'ASU vers la LSU.

¹³⁵ Version demo disponible en ligne : <http://www.uyghurdictionary.org/tools.asp> , version hors ligne sous forme de logiciel plug-in pour Microsoft Word: <http://oyghan.com/OTB/index.html>

¹³⁶ Cet outil a été mis en ouvre avec Visual Basic, sous forme d'un Macro.



« ئويغان » يۇمشاق دېتاللىرىنى köp rexmet ishletkingizge

Figure 12. vue après la conversion de l'ASU vers la LSU.

La version en ligne de cet outil permet aux internautes de : a) saisir les lettres ouïghoures en utilisant la méthode d'entrée au niveau du navigateur (intégrée dans l'outil) ; b) convertir des textes entre LSU, ASU et CSU ; c) normaliser en Unicode des textes écrits en utilisant les polices non-Unicode du Groupe de Recherche 863 de l'Université du Xinjiang:

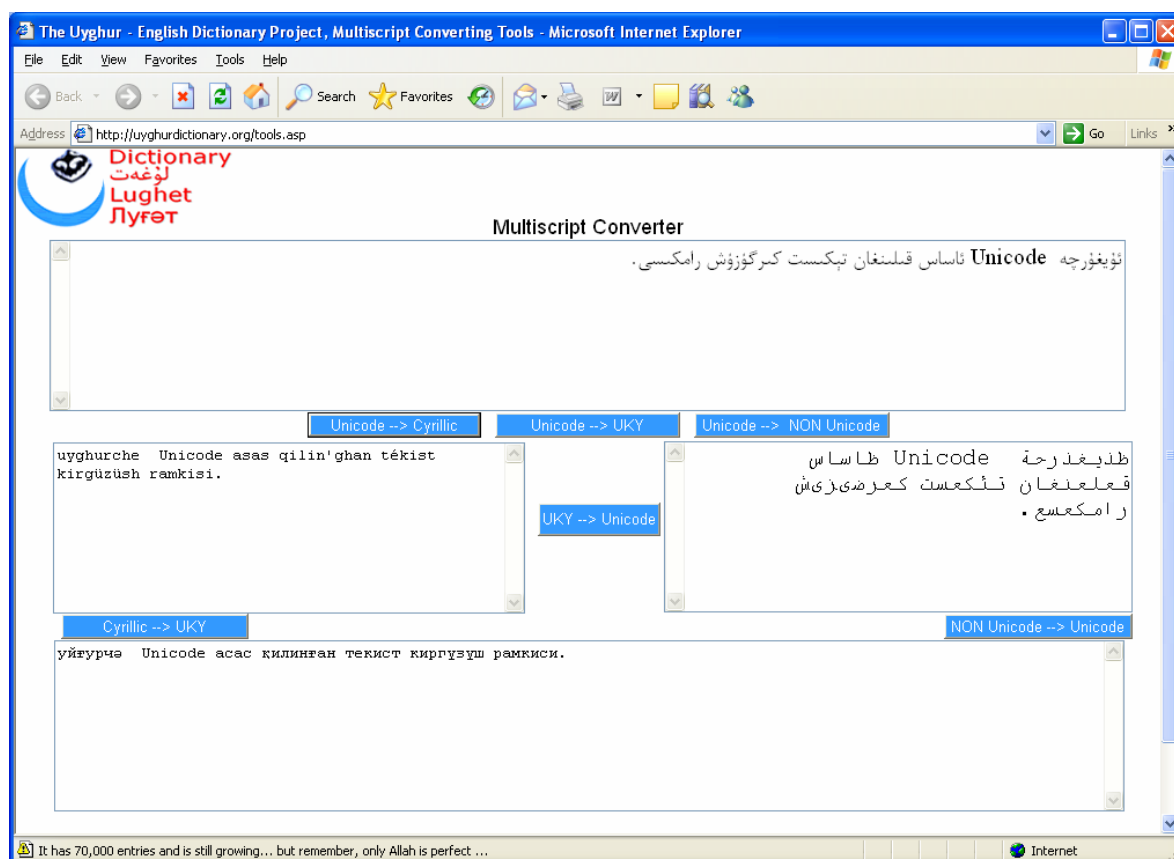


Figure 13. Convertisseur multi-écriture (en ligne) entre ASU $\leftrightarrow$ LSU $\leftrightarrow$ CSU $\leftrightarrow$ ASU non Unicode

Les convertisseurs multi-écritures présentés ci-dessus ne sont pas de nature linguistique. Il ne faut pas les confondre avec des outils de traduction automatique. Ils sont donc des outils de translittération, conçus pour une langue comme l'ouïghour qui utilise trois systèmes d'écritures et ils se contentent de convertir un texte écrit dans un alphabet

vers un autre. Cependant, cette méthode peut être appliquée à n'importe quelle langue turque qui présente la même caractéristique de polygraphie que l'ouïghour.

3.9 Evaluation et développement d'un dictionnaire en ligne

Afin de mettre en pratique l'ensemble de technique présentée dans ce chapitre, nous avons mis en ouvre un dictionnaire ouïghour – anglais en ligne, en utilisant 70,120 entrées préliminaires du dictionnaire ouïghour – anglais compilées depuis 22 ans par Dr. Jean R. Duval. Nous exposons en les points importants dans les sections suivantes.

3.9.1 Affichage des lettres ouïghoures

Les lettres ouïghoures sont affichées en utilisant une des polices Unicode les plus utilisées parmi les internautes ouïghours – UKIJ Tuz Tom¹³⁷. Prenant en compte des utilisateurs qui ne savent pas comment installer des polices, nous avons appliqué la méthode d'incorporation des polices (voir 4.6). Le téléchargement de la police indiqué ci-dessous est toutefois conseillé comme car 1) les polices incorporées ne sont compatible qu'avec le navigateur Internet Explorer ; 2) l'installation des polices peut accélérer l'ouverture des pages web puisque les navigateurs n'ont plus besoin de télécharger des polices incorporées.

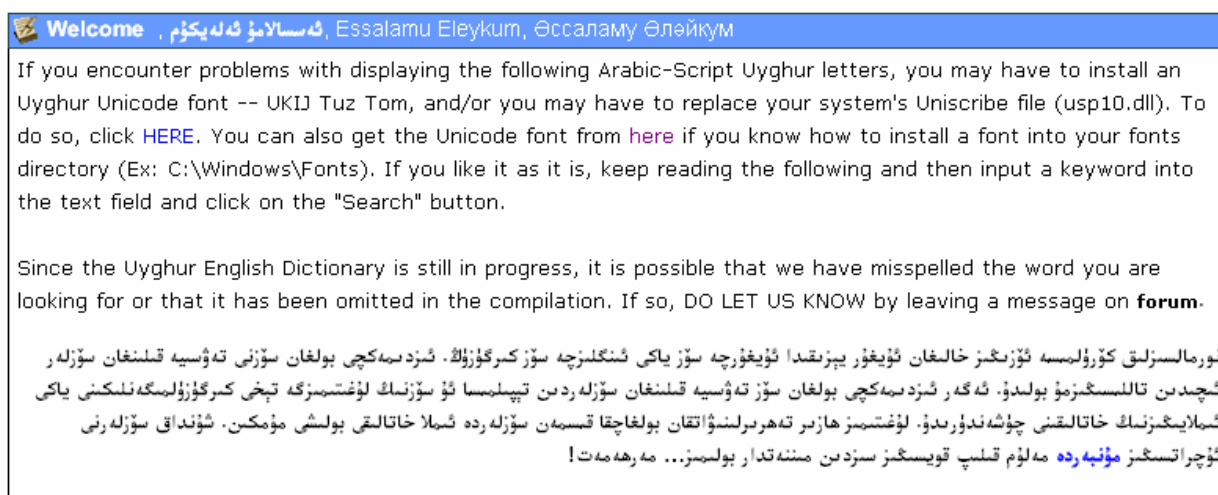


Figure 14. consignes de téléchargement de police pour ceux qui ont des problèmes d'affichage

3.9.2 Interface

¹³⁷ Cette police avait été développée par deux membres de l'Association Informatique Ouïghoure qui se sont inspirés de la première police Unicode, mentionné dans le chapitre 4.5.

Pour assister les utilisateurs du dictionnaire, l'interface et les messages d'instruction sont présentés en deux langues (anglais et ouïghour) et en trois écritures pour l'ouïghour. La boîte de texte s'aligne à droite lorsqu'on sélectionne la recherche par ASU (ئۇيغۇرچە), ou à gauche dans les autres cas pour s'adapter à la règle de direction de l'écriture. Cette fonctionnalité a été implémentée grâce aux attributs HTML de boîte de texte relatifs au contrôle de direction tels «dir= rtl, align=right ».

Figure 15. Interface de recherche

3.9.3 Saisie des mots à rechercher

La saisie des lettres en ASU est assurée par la méthode d'entrée au niveau de navigateur mentionnée dans la section 4.7 de ce chapitre. Cette méthode d'entrée se déclenche lorsque l'utilisateur choisit l'option de recherche par l'ASU et elle cède le contrôle de saisie à la méthode d'entrée ouïghoure au niveau système si l'utilisateur en est équipé. A partir du moment où il y a deux ou plus de deux lettres saisies dans la boîte de texte de recherche, le moteur de recherche interne enclenche un procédé et propose en temps réel les mots candidats (en ASU, LSU ou CSU selon l'écriture de saisie) qui commencent par les lettres saisies. L'utilisateur peut ensuite choisir l'un des mots proposés à l'aide la souris ou à l'aide des touches fléchées du clavier.

Figure 16. Suggestion automatique des mots candidats

Cette fonctionnalité est inspirée de celle qui a été développée par Google et implémentée en utilisant la technologie ASP.NET AJAX.

3.9.4 Recherche et présentation du résultat

Nous avons proposé quatre options de recherche, à savoir : exacte, commence par, terminé par, contient. L'option par défaut — recherche exacte, renvoie à l'entrée qui, à l'intérieur d'un champ d'une table de la base de données du dictionnaire, correspond exactement à la valeur de recherche précisée. Les autres options renvoient à toutes les entrées qui contiennent, commencent ou se terminent par la valeur de recherche. La recherche du mot saisi ou sélectionné commence par un clic sur le bouton « *search* ». Quand le mot est retrouvé dans le dictionnaire, le résultat de recherche est présenté dans les formes suivantes en tenant compte du nombre de résultats pertinents.

Figure 17. présentation du résultat de recherche : option de recherche → recherche exacte

Start With

ئالما

☐ ئۇيغۇرچە
☐ Uyghurche
☐ English
☒ ئۇيغۇرچە

ئالما	, alma, алма
ئالما قورۇمىسى	, alma qorumisi, алма қорумиси
ئالما قەقى	, alma qeqi, алма қеқи
dried apple pieces.	
ئالما قىيامى	, alma qiyami, алма қийами
ئالما مۇرابباسى	, alma murabbasi, алма мураббаси
ئالما ئۇرۇقدىشى	, alma uruqdish, алма уруқдиши
ئالما ئۇرۇقى رەڭ	, alma uruqi reng, алма уруқи рەң
ئالماس	, almas, алмас
ئالماس پارقىراقلىقى	, almas parqiraqliqi, алмас парқирақлиқи
ئالمаш	, almash, алмаш

Figure 18. présentation du résultat de recherche : option de recherche → commence par

Le résultat de recherche est vide si : a) le mot clé recherché n'existe pas dans le dictionnaire ; b) l'utilisateur atteint le nombre maximum de mots autorisés à la consultation quotidienne¹³⁸ ; c) le mot clé n'est pas saisi correctement. Dans le premier cas, le mot non-retrouvé est ajouté automatiquement à la liste des futures entrées potentielles ; dans le deuxième cas l'utilisateur reçoit un message d'alerte ; et dans le dernier cas le moteur de recherche fait une suggestion de correction d'orthographe. Nous présentons les détails de cette fonctionnalité dans la section suivante.

3.9.5 Correction d'orthographe

Après avoir étudié les mots recensés dans la liste des mots non-trouvés pendant le premier trimestre après le lancement du dictionnaire en ligne, nous avons constaté que la grande majorité des fautes réside dans les problèmes suivants (par ordre de fréquence) :

- Influence de la langue parlée: en raison des dialectes, de la consonance étrangère des mots empruntés ou des multiples réformes d'orthographe, les gens ont tendance à écrire comme ils prononcent. Ex :

Correct	Incorrect	Correct (emprunt)	Incorrect	Correct	Incorrect (dialecte)
<i>Fontan,</i>	<i>Pontan</i>	<i>Radio</i>	<i>radiyo</i>	<i>Yiltiz</i>	<i>Zhiltiz</i>

¹³⁸ Afin de lutter contre la piraterie, nous avons limité le nombre de consultations à partir de la même adresse IP à 100 par jour.

<i>fransiye</i>	<i>piransiye, pransiye, firansiye</i>	<i>Kompyutér</i>	<i>kompitur, kompitor, kompiyotir</i>	<i>Qir</i>	<i>Qizh</i>
<i>Pajie</i>	<i>paje, faje, fajie</i>	<i>Shtat</i>	<i>shitat, ishtat</i>	<i>gösh</i>	<i>Gosh</i>
<i>Kitab</i>	<i>Kitap</i>	<i>Printér</i>	<i>pirintir</i>	<i>xoraz</i>	<i>Ghoraz</i>
<i>Ijad</i>	<i>Ijat</i>	<i>Drama</i>	<i>dirama, diramma, daramma</i>	<i>kirmek, kiring</i>	<i>Chirmek, chiring</i>
<i>Gül</i>	<i>Gul</i>	<i>Chashka</i>	<i>chachka, chechke</i>	<i>deptuq</i>	<i>Deptük</i>
<i>Ördek</i>	<i>Ordek, ödek,</i>	<i>Zhurnal</i>	<i>jornal, jurnal</i>	<i>qarimaq, qara</i>	<i>Qaylimaq, qala, qayla</i>
<i>Ereb</i>	<i>Erep</i>	<i>Stolba</i>	<i>Istolba</i>	<i>qochqar</i>	<i>Qoshqa, qochqa</i>
<i>Janbaz</i>	<i>Jambaz</i>	<i>Awwal</i>	<i>awal</i>	<i>chipta</i>	<i>Chupta</i>
<i>Kélin</i>	<i>Kilin</i>	<i>Téléfon</i>	<i>Tilpun, tilipun</i>	<i>toxu</i>	<i>Toxuy, toxa</i>
<i>Lughet</i>	<i>Loghet</i>	<i>Layihe</i>	<i>laye, lahiye</i>	<i>ténchliq</i>	<i>Chizhliq</i>
<i>qilmaq</i>	<i>qimmaq</i>	<i>grammatika</i>	<i>girammatika, girammatka</i>	<i>déyelmeymen</i>	<i>Delemeymen</i>
<i>musht</i>	<i>Mush</i>	<i>pénsiye</i>	<i>pissiye, pinsiye</i>	<i>chömüch</i>	<i>Chömche</i>

Table 9. Exemple des fautes d'orthographe

Les exemples ci-dessus couvrent des fautes telles que la confusion –p/-f, -ap/-at/-et (-ab/-ad/-ed), le changement de voyelles (ö, ü, u, o, i, é), le remplacement de « -n-, -l- » par « -m- », l'« assimilation » de syllabes étrangères à des syllabes ouïghoures par rajout de voyelle (dans la plupart des cas : « i ») ou la modification d'une voyelle. Elision, insertion ou changement d'une ou plusieurs lettres dans le cas des erreurs dues à l'influence dialectale sont aussi à prendre en compte. Mots non trouvés sont analysés par un procédé de

suggestion dans lequel les cas mentionnés ci-avant sont pris en compte afin de suggérer des mots correctement orthographiés.

Certaines erreurs sont dues à une mauvaise maîtrise de la disposition du clavier (norme régionale). L'utilisateur ne regarde pas l'écran quand il saisit mais se concentre sur le clavier. Ex :

Symptôme	incorrect	correct	Cause
Mauvaise maîtrise de la disposition du clavier	كاندهك، نهچچه	قانداق، تهچچه	L'utilisateur réfléchit en Pinyin ¹³⁹ et appuie sur les touches marquées "k, a, n, d, a, k", « n, a, q, q, a » en espérant pouvoir écrire « نهچچه », « قانداق ».
Oubli ou non fonctionnement de la touche "shift"	ئىقاد، بوتهن	ئىجاد، خوتهن	« ژ، ف، گ، خ، ج، و » nécessitent la combinaison de "shift" avec les touches "k, j, h, g, f, d" pau moment de la saisie pour ne pas être réalisés en « د، ا، ه، ی، ق، ل »
Problème de touches	ياتسۇق، ئاتماقق	ياتسۇق، ئاتماق	En raison de la configuration courte du «délai de répétition » (repeat delay) ou de la durée excessive de la pression sur les touches, insertion d'une ou plusieurs lettres non souhaitées (apparition successive de quatre t, et de trois q)
«n'importe quoi »	ادىسكالىسى، اڭپۇرۇۋېرېۋ	pas de correction	L'utilisateur joue avec les touches du clavier et fait des fautes volontaires en tapant des «mots», du genre « djsfjdsklaf, ruiewoqrewoq »

Parmi les mots non trouvés, il figure aussi des mots suffixés qui ne font pas partie du dictionnaire. Ceci n'est pas pris en compte par le procédé de suggestion d'orthographe, mais nous considérons qu'il pourra être traité en utilisant les méthodes mentionnées dans le

¹³⁹ Si l'utilisateur utilise la disposition du clavier régional, basée sur le Pinyin. En cas d'utilisation de la disposition du clavier basée sur LSU, ça pourrait être «هوزا» (touches : h, o, r, a, z) au lieu de «خورا» (touches: x, o, r, a, z) si l'utilisateur ne se souvient pas que «x» est prononcé comme son équivalent en LSU – «خ».

chapitre 6 et 8. Mots non trouvés sont présentés dans la forme suivante, avec des mots suggérés qui existent dans le dictionnaire. Utilisateur peut ensuite cliquer sur un des mots suggérés pour obtenir son explication.

Figure 19. Suggestion de correction d'orthographe

3.9.6 Outil de recherche pour les webmasters

Afin de préconiser ce dictionnaire en ligne et d'attirer des internautes sur notre site, nous avons développé un outil qui peut être intégré dans les sites web des internautes. Il est possible de l'intégrer dans un des modes (mode complet, mode simple, mode en LSU/ASU/CSU) en suivant des étapes ci-dessous :

Figure 20. Recherché par ASU

Figure 21. Recherché par LSU

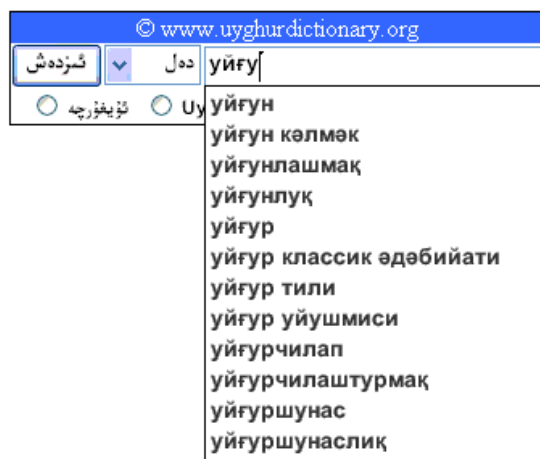


Figure 22. Recherche par CSU

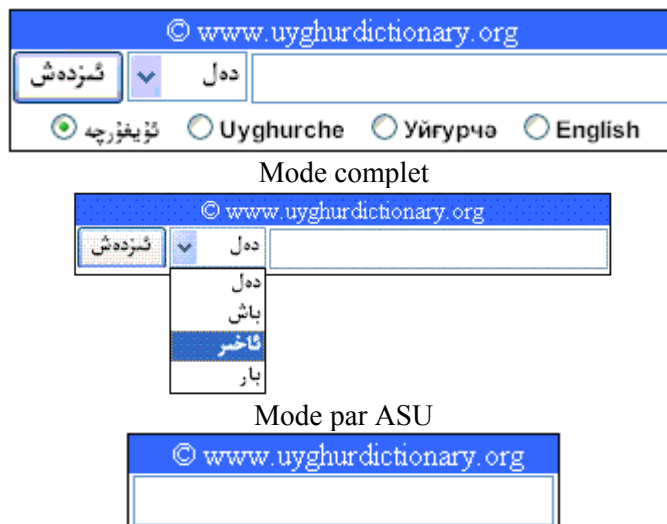


Figure 23
Mode simple

- Ajouter le code ci-dessous entre les balises <head>...</head> ;

```
<link type="text/css" href="UyghurDictionaryOrgSearch.css" rel="stylesheet"/>
<script src="UyghurDictionaryOrgSearch.js"></script>
<script src="YulghunCombedit.js"></script>
```

- Ajouter le code ci-dessous à la place de la balise <body> ;

```
<body
                                onload="document.getElementById('UyghurDictionaryOrg').focus();"
onClick="UyghurDictionaryOrgHide();">
```

- Ajouter le code ci-dessous là où on veut placer le gadget du dictionnaire ouïghour-anglais.

```

<div id="UyghurDictionaryOrgMain">
  <table width="323" id="UyghurDictionaryMainTable" cellpadding="0" cellspacing="0">
    <form
      name="UyghurDictionaryOrgSearchForm"
      method="get"
      action="http://www.uyghurdictionary.org/ug-arab/default.aspx" target="_blank">
      <input type="hidden" name="dictype" id="dictype" value="ue">
      <tr>
        <td colspan="3" class="UyghurDictionaryOrg_Copyright_Note">
          © www.uyghurdictionary.org</td>
        </tr>
        <tr>
          <td width="54"><input type="submit" name="UyghurDictionaryOrgSubmit"
            class="UyghurDictionaryOrgSubmit" value="ئىزدەش"></td>
          <td><span class="UyghurDictionaryLanguageOptions">
            <select name="so" id="so">
              <option value="ex" selected>ئەرەب</option>
              <option value="sw">باش</option>
              <option value="ew">ئاخىر</option>
              <option value="in">بار</option>
            </select>
          </span> </td>
          <td width="206">
            <input name="q" type="text" id="UyghurDictionaryOrg" maxlength="80"
              onKeyPress="addchar(this,event); UyghurDictionaryOrgStart();" dir="rtl"
              onmouseover="this.focus();"></td>
        </tr>
        <tr>
          <td colspan="3" height="19" class="UyghurDictionaryLanguageOptions"
            align="center">
            <label for="Input_UEY"><input name="lang" type="radio"
              id="Input_UEY" value="UEY" checked onChange="return oc('ue');">ئۇيغۇرچە </label>
            <label for="Input_ULY"><input name="lang" type="radio" value="ULY"
              id="Input_ULY" onChange="return oc('eu');">Uyghurche </label>
            <label for="Input_USY"><input type="radio" name="lang" value="USY"
              id="Input_USY" onChange="return oc('eu');">Уйғурчә </label>
            <label for="Input_EN"><input type="radio" name="lang" value="EN"
              id="Input_EN" onChange="return oc('eu');">English</label>
          </td>
        </tr>
      </form>
    </table>
  </div>
  <div id="UyghurDictionaryOrgResult" dir="rtl"></div>

```

Lorsque l'utilisateur clic sur un des mots suggéré, il est renvoyé vers le site de dictionnaire pour voir le résultat de recherche. Ci-dessous est un rapport du nombre de visiteur depuis la création du dictionnaire ouïghour-anglais :

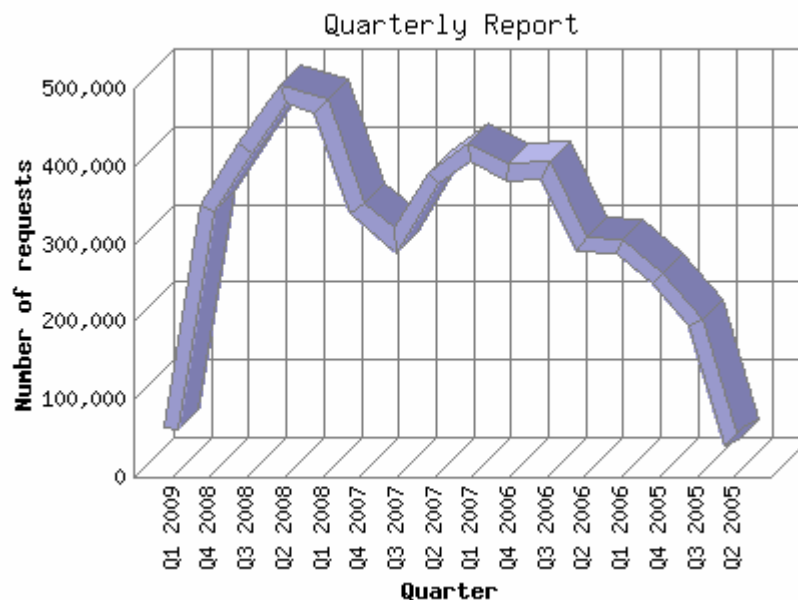


Figure 24. trafic trimestriel du site du dictionnaire (image présentée par l'outil de statistique -- *Report Magic*, fourni par l'hébergeur sous forme d'un outil interne du site)

Chapitre 5. Suffixes verbaux du ouïghour

Ce chapitre explique la définition des règles de suffixation de l'ouïghour moderne, une langue turque agglutinante marquée par une forte harmonie tant vocalique que consonantique. Ce chapitre commente les variations qui en découlent et s'attache également à la correspondance entre le niveau de surface et le niveau lexical avant de se concentrer sur la morphologie et l'ordre de succession des suffixes verbaux. Des règles y sont développées pour permettre l'extraction de données informatiques sur les verbes ouïghours et créer une base pour le développement d'un système complet qui couvrirait toutes les entités lexicales de la langue ouïghoure. L'objectif final de cette recherche en cours est de proposer une approche linguistique plutôt qu'une approche traditionnelle par élaboration de corpus (*corpusing*) pour le traitement automatique de la langue ouïghoure (TALO). Celle-ci pourra, à son tour, être appliquée à la mise au point d'un logiciel de correcteur orthographique qui utiliserait un corpus lexical minimal complété par des règles complètes de suffixation pour usage dans le traitement de texte et la reconnaissance optique de textes en ouïghour. Ce logiciel pourrait être adapté plus tard pour des langues similaires telles que l'ouzbek, le kazakh ou le kirghiz.

5.1 Introduction

L'ouïghour moderne écrit varie du ouïghour parlé, puisqu'il est une représentation phonémique plutôt que phonétique de la langue parlée standard¹⁴⁰. La distance qui existe entre la forme écrite et la prononciation réelle de la langue¹⁴¹, ainsi que l'existence de dialectes causent de fréquents problèmes d'orthographe. Plusieurs réformes de l'alphabet et de l'orthographe ont été tentées au cours de la deuxième moitié du vingtième siècle mais, le plus souvent, plutôt que résoudre le problème¹⁴² elles l'ont aggravé. Un certain nombre

¹⁴⁰ Pour une description détaillée de la prononciation et de ses complexités, voir R.F.Hahn *Spoken Uyghur*.

¹⁴¹ par exemple:: *aptobus* généralement prononcé *awtowuz*.

¹⁴² Voir J.R. Duval, *Modern Uyghur: a Historical Perspective*.

d'incohérences affectant particulièrement les lettres *b* et *d* en position finale¹⁴³ est aussi une source de mauvais orthographe. Aussi arbitraires et incohérentes soient-elles, les règles d'orthographe de l'ouïghour moderne simplifient la tâche du correcteur orthographique dans le sens où elles offrent une alternative systématisée aux variantes¹⁴⁴ parlées et permettent de traiter de manière assez efficace les changements multiples qui affectent consonnes et voyelles dans les divers environnements phonétiques de la langue parlée. Dans cette étude nous avons choisi de suivre l'orthographe de l'ouïghour moderne telle que décrite dans le *Hazirqi Zaman Uyghur Edebiy Tilining Teleppuz Lughiti* (Dictionnaire de prononciation de la langue littéraire ouïghoure moderne) de 1988, comme nous l'avons fait dans notre dictionnaire ouïghour-anglais en ligne¹⁴⁵, plutôt que celle du guide officiel d'orthographe de 1985 *Hazirqi Zaman Uyghur Edebiy Tilining Imla Lughiti* (Dictionnaire d'orthographe de la langue littéraire ouïghoure moderne), qui n'est pas à jour. Les variances d'orthographe historiques et dialectales devront être prises en compte plus tard, lorsque des correcteurs orthographiques seront développés pour la reconnaissance optique de textes non-contemporains.

Malgré un nombre de projets rapidement abandonnés qui ont été lancés dans le cadre de recherche de traduction chinois-ouïghour dans la RAOX¹⁴⁶, TALO reste généralement un territoire vierge. De multiples recherches académiques faites en Turquie sur NLP turc depuis les années 90, notamment par A.Solak et K.Oflazer¹⁴⁷, ne sont pas directement applicables à l'ouïghour en raison des défis spécifiques que celui-ci présente. L'ouïghour partage un nombre de caractéristiques avec les autres langues turques, autant en termes de structure (agglutination) que de phonétique (harmonie vocalique).

¹⁴³ Ex. : *seweb*, *kitab* mais *mektep* et *kutupxana*; tous en relation avec la racine arabe *K-T-B*; *zawut* (russe : *zavod*), *meqset* (arabe : *meqsed*) but *dad* (persan : *dad*).

¹⁴⁴ Ex. : l'adoucissement dans la langue parlée de quelques consonnes finales (*k* → *g*, *q* → *gh*, *-b* / *-p* → *w*) devant les suffixes commençant par un *I* (par exemple: *mektep* → *mektiwi*) n'est plus reflété, sauf dans quelques formes verbales où l'adoucissement a été depuis longtemps avalisé par l'usage (par exemple: **kétip idim* → *kétiwidim*).

¹⁴⁵ Voir <http://www.uyghurdictionary.org>

¹⁴⁶ Voir <http://scholar.ilib.cn/Abstract.aspx?A=zwxxxb200505011>.

¹⁴⁷ Voir Ayşin Solak, Kemal Oflazer, 1993, "Design and Implementation of a spelling checker for Turkish", Kemal Oflazer, 1994, Two-level Description of Turkish Morphology, et Cüneyd Tantıg et Esref Adalı et Kemal Oflazer, 2006, Computer Analysis of the Turkmen Language Morphology.

L'exemple suivant démontre la nature agglutinante de la formation des mots en ouïghour. Le « mot » *Küchlendürelmeywatqanliringlarningkidinmu?* Correspond à une phrase française complète, «Sont-ils parmi ceux que tu es incapable de renforcer ? ». La racine dans ce mot est « *küch* » (force, pouvoir) qui contient une voyelle antérieure arrondie « *ü* ». La suite de suffixes ajoutés à la racine se présente comme *le+n+dür+el+mey+wat+qan+lir+ing+lar+ning +ki+din+mu* (le symbole « + » marque une frontière de suffixe). Chaque suffixe modifie l'information sémantique de la racine. Pour être plus précis, *+le+* crée une forme verbale d'harmonie vocalique postérieure, *+n+* donne un verbe réfléchi, *+dür+* génère la forme factitive d'une racine d'harmonie vocalique antérieure après une consonne sonore, *+el+* est la forme du potentiel des verbes d'harmonie vocalique antérieure et est suivie du suffixe négatif d'harmonie vocalique antérieure *+mey+* ; *+wat+* est un marqueur aspectuel (temps continu) qui joue un rôle de transformateur¹⁴⁸ vers l'harmonie vocalique postérieure ; il est suivi du marqueur de participe passé de participe passé de voyelle postérieure *+qan+* dont la consonne initiale sourde correspond à la consonne finale sourde du marqueur précédent ; *+lir+* est la marque du pluriel de voyelle antérieure affaiblie suivie de *+ing+*, la seconde personne singulier de voyelle postérieure ; *+lar+* est la marque du pluriel de voyelle postérieure; *+ning+* est la marque vocaliquement neutre du génitif ; *+ki+* est un relatif vocaliquement neutre ; *+din+* est la marque vocaliquement neutre de l'ablatif après une voyelle ou une consonne sonore et, finalement, *+mu+* est la marque vocaliquement neutre de l'interrogatif.

Dans des langues agglutinantes telles que l'ouïghour et d'autres langues turques, plusieurs suffixes peuvent être fixés à une racine et les frontières morphémiques dépendent du contexte morphologique et phonologique. Mais l'ouïghour possède aussi des fonctions morphologiques complexes uniques que nous allons voir dans les prochains paragraphes. Les traits communs à toutes les langues turques aussi bien que les traits spécifiques de l'ouïghour demandent une attention particulière lors de l'identification et la mise en ordre de ses nombreux suffixes et dans la détermination des interconnexions. Cette complexité fait que le TALO est très difficile à implémenter, particulièrement lors de l'analyse linguistique. Des erreurs dans l'harmonie des consonnes et des voyelles ainsi qu'un ordre incorrect des suffixes causent des fautes d'orthographe. Il est évident que collecter toutes

¹⁴⁸ Voir 5.5

les formes suffixées possibles des racines dans un corpus serait quasi-impossible et pourrait causer d'importantes et inutiles redondances d'information. Afin de clarifier le processus de suffixation dans l'ouïghour écrit, il faut d'abord définir les règles phonétiques qui l'affectent. Dans les sections suivantes nous allons les décrire telles qu'elles s'appliquent à notre projet de recherche en cours sur la définition des règles de suffixation de l'ouïghour.

5.2 Consonnes

Les consonnes ouïghoures sont théoriquement divisés en consonnes sourdes (ou dures), C_h {ch, f, h¹⁴⁹, k, p, q, s, sh, t, x} et en consonnes sonores (ou douces) C_s (b, d, g, gh, j, l, m, n, ng, r, w, y, z, zh). Mais en position finale dans les radicaux, les lettres *b*, *d* et *g* sont respectivement traités comme *p*, *t* et *k* dans beaucoup de contextes. Ils seront nommés ci-après B (labiales b ou p), D (dentales d ou t) et K (laryngo-vélaires g, gh, k ou q) afin de prendre en compte ces fluctuations et le fait qu'au niveau lexical un certain nombre de suffixes commence par une consonne qui, d'après le contexte phonétique, peut être réalisé comme d/t ou g/k/gh/q. Le changement de consonne passe dans plusieurs contextes¹⁵⁰ : (1) lorsqu'un *i* ou un suffixe commençant par *i* est ajouté à un *p* final dans certains mots, la valeur phonétique, reflétée par l'orthographe, devient *b*¹⁵¹. $B + 0 \rightarrow \{b/p\}$ (par exemple: {kitap}), $B + i \rightarrow \{bi\}$ ¹⁵² (par exemple: {kitab*i*}); (2) *-g* et *-gh* finaux « s'endurcissent » (c'est à dire qu'ils sont prononcés comme leurs équivalents non prononcés *-k* et *-q*) lorsque suivis d'un suffixe commençant pas K, même si ce changement ne se reflète pas dans l'écriture. L'initiale du suffixe s'endurcit aussi et ceci se reflète dans son orthographe (par exemple: tugh+ ghan $\rightarrow$ tughqan [pron: tuqqan], teg + gen $\rightarrow$ tegken [pron: tekken], tagh + gha $\rightarrow$ taghqa [pron: taqqan]).

¹⁴⁹ La lettre h en position finale est souvent abandonnée dans la langue parlée, laissant la voyelle précédente dans une position finale (par exemple : . *aramgah*, épelé *aramgahqa* au datif mais souvent adouci en *aramgâgha* dans la langue parlée). Ce phénomène a amené certains phonéticiens ouïghours à considérer à tort h comme une consonne douce.

¹⁵⁰ N'est pas respecté dans le dictionnaire de prononciation de 1988 *op.cit.*

¹⁵¹ Souvent prononcé comme un w. Cet adoucissement n'arrive jamais dans des mots où un p final est dérivé d'un *f* étymologique, tels *terep* [$\leftarrow$ *teref] ou *sinip* [$\leftarrow$ *sinif].

¹⁵² A noter que la prononciation hésite entre {b} et {w}.

5.3 Voyelles

La différenciation des voyelles se fait sur la base de la position de leur émission dans la bouche (voyelles d'avant ou antérieures par opposition à voyelles d'arrière ou postérieures) le degré d'ouverture des lèvres (voyelles arrondies par opposition à voyelles non arrondies) et la hauteur de la langue (voyelles hautes par opposition à voyelles basses) lors de leur production..

	Non arrondie (V _u)		arrondie (V _r)	
	antérieure(V _f)	postérieure(V _b)	antérieure (V _f)	postérieure (V _b)
Haute	i, é	i, ¹⁵³ é	ö	O
Basse	e	a	ü	U

Table 5. voyelles

L'ouïghour écrit et l'ouïghour parlé ne différencient pas entre l'é et l'i¹⁵⁴ d'avant et l'é et l'i d'arrière, mais ces deux séries de voyelles impartissent ou transmettent un caractère antérieur ou postérieur suivant le cas aux suffixes qui les suivent. Dans cette étude, nous avons établi de manière artificielle une différence entre une valeur d'avant ou d'arrière de ces deux voyelles.. On devrait aussi noter que les deux lettres peuvent être présentes dans la forme lexicale d'origine d'un mot ou suffixe (par exemple: *kiyim*, *uniwérsité*, *téléfon*, *téz*) ou résulter d'un phénomène connu sous le nom d'affaiblissement vocalique (balilar ← bala+lar, kélidu ← kel+i+du).

L'affaiblissement vocalique affecte les voyelles *a* et *e* dans la syllabe finale des mots plurisyllabiques ou dans la syllabe unique d'un mot ou suffixe monosyllabique lorsqu'elle est suivie d'une voyelle dont elles ne sont séparées que par une seule consonne. Dans ce cas, ce sont des allomorphes affaiblies de *a* (V_{bw}) et *e* (V_{fw}). Par exemple: kel- → kéling, al-

¹⁵³ Certains linguistes décrivent i et é comme des voyelles médianes, du fait qu'elles peuvent être aussi bien antérieures que postérieures.

¹⁵⁴ Les différences de prononciation, spécialement dans le cas de i, sont dues à l'environnement de la consonne et non à la valeur postérieure ou antérieure du mot dans lequel elles se trouvent (voir *Spoken Ouïghour*, Reinhard F. Hahn, 1991).

→ éling; sözle- → sözligin, atla- → atlighin. La règle générale établie par les grammairiens¹⁵⁵ ouïghours est que l'affaiblissement de voyelle entraîne la transformation d'un *a* ou *e* en *é* dans des mots monosyllabiques (par exemple: al- → éli, kel- → kéli-) et en *i* dans des mots polysyllabiques ainsi que dans la plupart des suffixes¹⁵⁶ (par exemple: terep → teripi, bala → balisi, kel-se → kelsimu). Notre analyse nous a permis de raffiner cette règle générale : dans des noms communs, seuls *a* long et *e* long (ci après â et ê) échappent à l'affaiblissement (par exemple: kitâP → kitâbi, weqê → weqêsi); les radicaux verbaux monosyllabiques en *a* sont affectés par l'affaiblissement uniquement lorsqu'ils sont suivis d'un *i* euphonique (voir paragraphe 6.2) et aucunement dans d'autres cas (par exemple: élip, élish, but alar, alimen, alidighan).

Dans certains mots à deux syllabes, la deuxième voyelle (en général une voyelle aigüe *i* ou *u*) n'est rien de plus qu'un soutien euphonique entre deux consonnes, et disparaît lorsqu'un suffixe commençant par une voyelle est ajoutée au mot (par exemple: burun + i → burni, singil + i → singlim, isim + i → ismi). Cette voyelle elliptique sera appelé ci-après \$V (par exemple: bur\$un). Parfois, la présence d'une voyelle elliptique causera l'affaiblissement de la voyelle dans la syllabe précédente (par exemple: aghzi – i → éghiz) compliquant ainsi la recherche du radical d'un mot.

5.4 L'harmonie des consonnes

Certains suffixes commencent par une dentale ((D:{d/t}¹⁵⁷) ou une laryngo-vélaire (K: {g/k/gh/q}). Les règles suivantes s'appliquent¹⁵⁸ au choix de la consonne initiale dans un suffixe, basée sur la lettre finale du mot tel qu'il est attaché :

(1) une consonne finale dure fait appel à une consonne dure: C_h__C_h. par exemple: kitaB + Din → kitaBDin: {kitabtin}, tok + Ka → tokKa: {tokqa}.

¹⁵⁵ Voir Kaşgarlı 1992, *Modern Uyğur Türkçesi Grameri*, Qazaq Penler Akademiyisi 1966, *Hazırqi Zaman Ouighour Tili, 2-qisim, Morfologiyeye we Sintaksis*, et Tömür 1987, *Hazırqi zaman Ouighour tili grammatikisi (morfologiyeye)*.

¹⁵⁶ Les suffixes verbaux évolués de verbes (-wet- [← yet-], -wer- [← ber-], -wal- [← al-] suivent la règle qui s'applique aux racines verbales d'une syllabe (par exemple: kéliwétip, kütüwérish, soruwélsh).

¹⁵⁷ Les formes de surface sont notées avec des accolades. La lettre majuscule A représente tant *a* que *e* ; K représente *g/k/gh/q*.

¹⁵⁸ Pour plus de détails à propos des abréviations, voir annexe 5.

(2) une consonne finale douce fait appel à une consonne douce: $C_s\_C_s$. par exemple: pul + Din $\rightarrow$ pulDin: {puldin}, nur + Ka $\rightarrow$ nurKa : {nurgha}.

(3) une voyelle finale fait appel à une consonne douce: V_C_s . par exemple: ana + Din $\rightarrow$ aniDin: {anidin}, jüme + Ke $\rightarrow$ jümeKe : {jümege}, yöle + Di $\rightarrow$ yöliDi: {yölidi}.

(4) dans une position initiale de suffixe, deux séries de consonnes lexicales sont sujettes au changement :

a) les consonnes dentales D^* : {d/t}

$C_h+D \rightarrow C_h+\{t\}$: gep+De $\rightarrow$ {gepte}, chaq+Di $\rightarrow$ {chaqti}

$C_s+D \rightarrow C_s+\{d\}$: teg+Di $\rightarrow$ {tegdi}, biz+De $\rightarrow$ {bizde}

$V+D \rightarrow V+\{d\}$: al+Di $\rightarrow$ {aldi}, qur + Da $\rightarrow$ {qurda}

b) les consonnes laryngo-vélaires K^* {g/k/gh/q}

$V_f^*C_h+K^* \rightarrow V_f^*C_h+\{k-\}$: {ket-}+ K^* in $\rightarrow$ {ketkin}

$V_b^*C_h+K^* \rightarrow V_b^*C_h+\{q-\}$: {tap-}+ K^* in $\rightarrow$ {tapqin}

$V_f^*C_s+K^* \rightarrow V_f^*C_s+\{g-\}$: {kel-}+ K^* in $\rightarrow$ {kelgen}

$V_b^*C_s+K^* \rightarrow V_b^*C_s+\{gh-\}$: {qal-}+ K^* in $\rightarrow$ {qalghin}

$V_f+K^* \rightarrow V_f+\{g-\}$: {sözle-}+ K^* in $\rightarrow$ {sozligin}

$V_b+K^* \rightarrow V_b+\{gh-\}$: {qara-}+ K^* in $\rightarrow$ {qarighin}

(5) les consonnes laryngo-vélaires K dans le suffixe dubitatif K^*u {qu/ghu}. Etant donné que la voyelle est invariable, la consonne initiale de la particule finale dubitative est toujours K {q/gh}:

$V_f^*C_h+K^*u \rightarrow V_f^*C_h+\{qu\}$: {ketsek-}+ $K^*u \rightarrow$ {ketsekqu}

$V_b^*C_h+K^*u \rightarrow V_b^*C_h+\{qu\}$: {alsaq}+ $K^*u \rightarrow$ {alsaqqu}

$V_f^*C_s+K^*u \rightarrow V_f^*C_s+\{ghu\}$: {kelgen}+ $K^*u \rightarrow$ {kelgenghu}

$V_b^*C_s+K^*u \rightarrow V_b^*C_s+\{gh-\}$: {qalghan}+ $K^*u \rightarrow$ {qalghanghu}

$V_f+K^*u \rightarrow V_f+\{gh-\}$: {keldi}+ $K^*u \rightarrow$ {keldighu}

$V_b+K^*u \rightarrow V_b+\{gh-\}$: {aldi}+ $K^*u \rightarrow$ {aldighu}

(6) K terminant un suffixe en $SUF\ l^*I^*K$:

$V_{fu}^*C/0+l^*I^*K \rightarrow V_{fu}^*C/0+\{lik\}$ par exemple: Ürümchilik

$V_{bu}^*C/0+l^*I^*K \rightarrow V_{bu}^*C/0+\{liq\}$ par exemple: Turpan+liq

$V_{fr}^*C/0+l^*I^*K \rightarrow V_{fr}^*C/0+\{lük\}$ par exemple: күnlük

$V_{br}^*C/0+l^*I^*K \rightarrow V_{br}^*C/0+[luq]$ par exemple: Atushluq

(7) K en fin de suffixe de première personne pluriel $SUF *K V_f+ -K \rightarrow V_f+ \{-k\}$. Par exemple: $kelsek V_b+ -K \rightarrow V_b+ \{-q\}$: $alsaq, kelduq$

(8) assimilation de consonnes laryngo-velar (n'est pas pris en compte en ouïghour écrit) : $RAD(V*K_s) \_SUF(K*V*) \Rightarrow RAD(V*K_h) \_SUF(K*)$: $RAD(V_f*K(k)) \_SUF(K(k)*V_f*) / RAD(V_b*K(q)*V_b*)$ par exemple: $teg \rightarrow tekken, bagh \rightarrow baqqa, tugh \rightarrow tuqqan$.

5.5 Harmonie vocalique

Dans leur ensemble, les suffixes respectent l'harmonie vocalique lorsqu'ils sont attachés à une racine. Même dans des cas où ils ne sont pas affectés de manière morphologique, ils transmettent la valeur vocalique de la syllabe précédente. La dernière voyelle dans une racine fait appel en général à un suffixe qui appartient à la même classe vocalique. Certains suffixes sont aussi affectés par une harmonie à quatre sens. Il y a, par contre, quelques suffixes « forts » qui changent la valeur d'avant ou d'arrière de la chaîne de suffixes qui les suivent en fonction de leur propre valeur d'avant ou d'arrière. Finalement, quelques suffixes, que nous qualifierons de vocaliquement neutres, ne sont pas affectés par l'harmonie vocalique et n'en transmettent aucune. Ils se trouvent généralement retrouvés en position finale ou devant un auxiliaire.

Les règles ci-dessous s'appliquent lors de la sélection des suffixes :

(1) Harmonie à deux sens : une voyelle d'avant fait appel à une voyelle d'avant et une voyelle d'arrière fait appel à une voyelle d'arrière :

$*V_f* \_SUF(*A_f*)$: $\{e/\acute{e}\text{-antérieur}/i\text{-antérieur}/\acute{o}/\ddot{u}\} \_ *e*$. Par exemple: $k\acute{e}l + Ar \rightarrow \{k\acute{e}ler\}$, $kel + mAK \rightarrow \{kelmek\}$, $kel + KAn \rightarrow \{kelgen\}$, $\acute{o}y + lAr \rightarrow \{\acute{o}yler\}$.

$*V_b* \_SUF(*A_b*)$: $\{a, \acute{e}\text{-antérieur}, i\text{-antérieur}, o, u\} \_ *a*$ par exemple: $al + mAK \rightarrow \{almaq\}$, $\{b\acute{e}siq\} + mAK \rightarrow \{b\acute{e}siqmaq\}$, $qil + mA \rightarrow \{qilma\}$, $\{toy\} + KA \rightarrow \{toygha\}$, $ot + KA \rightarrow \{qa\}$.

(2) Harmonie à quatre sens: une voyelle non-arrondie fait appel à une voyelle non-arrondie et une voyelle arrondie fait appel à voyelle arrondie:

$*V_{fu}* \_SUF(*H_{fu}*)$: $\{e/\acute{e}\text{-antérieur}/i\text{-antérieur}\} \_ *i\text{-antérieur}*$

par exemple: $\{kel\} + Dim \rightarrow \{keldim\}$, $\{t\acute{e}rik\} + Di \rightarrow \{t\acute{e}rikti\}$, $\{tik\} + Kin \rightarrow \{tikkin\}$.

$*V_{bu}* \_SUF(*H_{bu}*)$: $\{a/\acute{e}\text{-postérieur}/i\text{-postérieur}\} \_ *i\text{-postérieur}*$

par exemple: al+Dim → aldim, tiq+Dim → tiqtim, béliq + im → béliqim

$*V_{fr}*\_\text{SUF}(*H_{fr}^*):*\{\ddot{o}/\ddot{u}\}*\_\text{ü}^*$

par exemple: küt + Düm → {küttüm}, {öl}+Düm → {öldüm}, {kör}+Küm → {körgüm}.

$*V_{br}*\_\text{SUF}(*H_{br}^*):*\{o/u\}*\_\text{u}^*$

par exemple: {tut}+Dum → {tuttum}, {bol}+Dum → {boldum}, {ot}+um → {otum}.

(3) Harmonie à deux sens dans les suffixes en voyelle basse arrondie : une voyelle antérieure fait appel à une voyelle basse antérieure arrondie {ü}, une voyelle postérieure fait appel à une voyelle basse postérieure arrondie {u}:

$*V_f*\_\text{SUF}(*H_{fr}^*):*\{e/\acute{e}\text{-antérieur}/i\text{-antérieur}/\ddot{o}/\ddot{u}\}*\_\text{ü}^*$ par exemple: kör+Küch → körgüch, kir + Kü → kirgü, küt + Kü → kütkü, söy + Kü → söygü.

$*V_b*\_\text{SUF}(*H_{br}^*):*\{a/\acute{e}\text{-postérieur}/i\text{-postérieur}/o/u\}*\_\text{u}^*$

par exemple: al+Ku → alghu, béqin + Ku → béqin'ghu, tut + Ku → tutqu, yoq + Ku → yoqqu.

(4) Harmonie à deux sens dans les suffixes en voyelle haute non-arrondie: une voyelle d'avant fait appel à une voyelle d'avant haute non arrondie (i) et une voyelle d'arrière fait appel à une voyelle d'arrière haute non arrondie {i}:

$*V_f*\_\text{SUF}(*H_{uf}^*):*\{e/\acute{e}\text{-antérieure}/i\text{-antérieure}/\ddot{o}/\ddot{u}\}*\_\text{i-antérieur}^*$

par exemple: kötürgü + chi, kelmek+chi, érit + Kin → éritkin, kör + Kin → körgin.

$*V_b*\_\text{SUF}(*H_{ub}^*):*\{a/\acute{e}\text{-postérieure}/i\text{-postérieure}/o/u\}*\_\text{i-postérieur}^*$

par exemple: ach + liq, bordaqchi+liq, al+Kin → alghin, uch + Di → uchti.

(5) Les suffixes transformateurs changent la valeur vocalique des suffixes qui les suivent :

$*V*\_\text{SUF}(*V_t^*)\_\text{SUF}*V(=V_t)^*$:

voyelle antérieure transformatrice {-wer-}, {-wet-}, {-[i]ken-}, {-ki-}, {-ger-}:

$*V*\_\text{SUF}(*V_{tf}^*)\_\text{SUF}(*V_f^*)$

par exemple:

{al=>él}+ i + wer + K*A*n → {éliwergen},

qala {qali} + wet + m*A*K → {qaliwetmek},

yol + D*A + {ki} + l*A*r → {yoldikiler},

{soda=>sodi}+ {ger} + l*A*r → {sodigerler},

{buz}+{u}+{p}+D*u {tu => ti}+{ken}+d*A → buzuptikende.

Voyelle postérieure transformatrice {-wal-}, {-wat-}, {-[i/y]dighan-}, {-nglar-}, {-dar-}, {-kar-}, {-zar-}:

$*V^* \text{__SUF}(*V_{tb}^*) \text{__SUF}(*V_b^*)$

par exemple:

$\{\text{kel} \Rightarrow \text{kél-}\} + i + \{-\text{wal-}\} + K^*A^*n \rightarrow \{\text{kéliwalghan}\},$

$\{\text{chüşh-}\} + i + \{-\text{wat}\} + K^*A^*n \rightarrow \{\text{chüshiwatqan}\}, \{\text{kör-}\} + i + \{\text{dighan}\} + l^*A^*r \rightarrow$

$\{\text{köridighanlar}\}, \{\text{üzüm}\} + \{-\text{zar}\} + l^*A^*r \rightarrow \{\text{üzümzarlar}\},$

$\{\text{emel}\} + \{-\text{dar}\} + l^*A^*r \rightarrow \{\text{emeldarlar}\}.$

(6) Suffixes vocaliquement neutres:

$*V^* \text{__SUF}(*V_{\text{neut}}^*)$, par exemple: men, sen, miz..., ghu/qu (dubitatif), -i- (auxiliaire du passé), -p, mu, mikin, -la (limitatif), -mish, -dur, -tur, -tu: ils peuvent se retrouver dans n'importe quel contexte vocalique et ne transmettent aucune valeur vocalique. Ils se trouvent généralement en position finale ou devant d'autres suffixes de voyelle neutre. Par exemple: alghanmen, kelgenmen, senghu?, ughu?, kéliw-i-di-m, qalghin-i-di-m, kéler-mikin, kelme-p-tu, alar-mikin, kéli-p-la, kel-di-mu.

5.6 Règles euphoniques dans la suffixation

1. Afin d'éviter les hiatus, une voyelle finale doit être séparée d'une voyelle initiale par une consonne (C_E) qui transmet la valeur vocale de la voyelle précédente. Dans le cas des verbes, cette consonne euphonique est un (y):

$V_f + C_E + 0^*V_f: \{\text{sözle} \Rightarrow \text{sözli}\} + \{y\} + \{-\text{el-}\} \rightarrow \{\text{sözliyel-}\}$

$V_b + C_E + 0^*V_b: \{\text{qara} \Rightarrow \text{qari}\} + \{y\} + \{-\text{al-}\} \rightarrow \{\text{qariyal-}\}.$

Pour les noms communs, cette consonne euphonique est généralement (s) mais aussi parfois (y) ou (r).

$V_b + C_E + 0^*V_b: \{\text{bala} \Rightarrow \text{bali}\} + \{s\} + \{i\} + K^*A \rightarrow \{\text{balisigha}\}, \{\text{balâ}\} + \{s\} + \{i\} + K^*A \rightarrow \{\text{balasigha}\}.$

$\{\text{bahâ}\} + l^*m \rightarrow \{\text{bahâyim}\} \text{ ou parfois } \{\text{bahârim}\}^{159}$

¹⁵⁹ Cette deuxième forme (r) est généralement considérée comme parlée et la langue standard n'accepte que la forme y.

2. Les suffixes composés d'une seule consonne (-n, -l, -sh, -p, -w; -m, -ng¹⁶⁰) nécessitent une voyelle euphonique lorsqu'ils suivent une consonne¹⁶¹. Cette voyelle euphonique I {i/i/ü/u} suit et transmet l'harmonie vocalique.

$V_{fr}^*C+I+C \rightarrow V_{fr}^*C+I_{fr}:\{köl\}+I+\{-m\} \rightarrow \{kölüm\},$

$V_{br}^*C+I+C \rightarrow V_{br}^*C+I_{br}:\{put\}+I+\{-ng\} \rightarrow \{putung\},$

$V_{fu}^*C+I+C \rightarrow V_{fu}^*C+I_{fu}:\{kel+kél-\}+I+\{-p\} \rightarrow \{kélip\}.$

Notez que I joue le rôle normal d'affaiblissement sur la voyelle précédente.

3. Les suffixes verbaux comportant une voyelle haute (D*H, l*H*K) subissent les mêmes changements que I, à l'exception du suffixe passé D*H+K de la première personne du pluriel, qui est toujours {duq/tuq}, et de la 3e personne et forme polie de la 2e personne du singulier, du passé D*H qui est toujours {di/ti}, quel que soit le contexte vocalique.

par exemple: {kör-}+D*H+{-m} $\rightarrow$ {kördüm},

mais :

{kör-}+D*H $\rightarrow$ {kördi},

{kör-}+D*H+-K $\rightarrow$ {körduq}

{kör-}+D*H+{-ngiz} $\rightarrow$ {kördingiz}

5.7 Morphologie

En ouïghour, plusieurs sortes d'affixes peuvent être ajoutées à une racine. Dans certains mots, on trouve des préfixes qui sont d'origine perse ou arabe (ex: bi+hajet, nim+jan, na+ilaj, kem + eqil). Mais les suffixes sont prédominants et omniprésents.

Il y a deux sortes de suffixes en ouïghour : 1) les suffixes de dérivation lexicale, qui modifient le sens ou parfois la catégorie de la racine et de créent de nouvelles unités lexicales (ex: *gül* – fleur, *gül+lük* $\rightarrow$ *güllük*–jardin); et 2) suffixes syntactiques, qui expliquent les relations syntaxiques ou fonctions variées du mot dans un contexte donné.

¹⁶⁰ Le suffixe -nglar est, en réalité, un suffixe composé : 0*C (-ng) + lar (suffixe transformateur postérieur pluriel). Il suit donc la même règle que -ng : *boldunglar*, *kélinglar*, *élinglar*, *körünglar*.

¹⁶¹ Dans ce système, nous avons choisi de traiter les suffixes commençant par une voyelle ou réduits à une seule consonne après une voyelle finale comme la combinaison d'une voyelle euphonique facultative et d'une consonne.

Les racines ouïghoures peuvent être classifiées en deux groupes majeurs : racines nominales et racines verbales. Le groupe nominal inclut les noms communs, les pronoms, les adjectifs et les nombres tandis que le groupe verbale inclut les verbes, qui ont la structure la plus complexe de suffixation. Les suffixes déverbaux changent les suites verbales (ie : racines verbales ou racines verbales étendues) en noms communs et les suffixes dénominiaux changent les noms communs en verbes. L'ordre de suffixation pour les groupes nominaux et verbaux peut être présenté comme suit :

Groupe nominal :

L'analyse de structure suivante reflète le cas le plus simple, c'est à dire le cas où il n'y a aucun suffixe dénominal qui change le nom commun d'origine en un verbe.

Racine étendue [racine+0/suffixe de dérivation] + 0/pluriel + 0/marqueur possessive + 0/cas¹⁶² + marqueur relatif + 0/[PLR][POS][CASE] + 0/[REL¹⁶³]... + 0/interrogatif¹⁶⁴ + marque de temps¹⁶⁵

Groupe verbal¹⁶⁶:

Les verbes, comme les noms communs, peuvent avoir une racine simple (primitive) et une racine étendue. Celle-ci est obtenue par l'affixation d'un nombre de suffixes dérivatifs (qui ajoutent un sens factitif, passif, potentiel, pluriatif ou coopératif, négatif ou interrogatif à la racine d'origine).

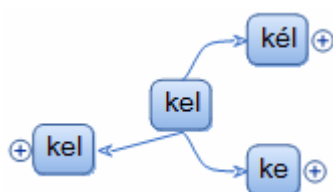


Figure 5.1 déformation de racine (vue sur l'explorateur de suffixes)

¹⁶² Ce sont l'ablatif, l'accusatif, le datif, le génitif et le locatif. Voir annexe 8 pour plus de détails.

¹⁶³ Les suffixes relatifs créent une boucle théoriquement infini, mais le nombre de boucles n'excède pas deux dans l'usage quotidien. Par exemple: *aili-di-ki-ler-ning-ki-din* (parmi ceux qui appartiennent à ceux de la famille), où *-ki* est le suffixe relatif.

¹⁶⁴ Variantes de l'interrogatif, voir annexe 8

¹⁶⁵ Les marqueurs de temps sont limités aux variantes de l'ancien verbe *er (être). Par exemple: *baliken* (← *bala+iken*) *yaxshidim* (← *yaxshi+idim*). Ils sont rajoutés à la suite des suffixes nominaux, le résultat de contractions qui sont communes dans la langue parlée et de plus en plus dans la langue écrite.

¹⁶⁶ Les suffixes actuels, à des niveaux différents, sont listés dans l'annexe 6.

Par l'ajout de marques syntactiques, plusieurs chemins différents peuvent être choisis qui mènent à plusieurs « produits finaux » — des verbes conjugués, des formes de compléments circonstanciels, ou des formes nominales :

Racine étendue [racine verbale¹⁶⁷ + 0/factitif + 0/reflexif/passif + 0/factitif + 0/pluratif + 0/factitif + 0/passif + 0/pluratif + 0/auxiliaire aspectuel + 0/factitif + 0/pluratif + 0/factitif + 0/passif + 0/négatif + 0/potentiel + 0/interrogatif] + *marqueurs syntactiques* [0/auxiliaire + 0/temps + 0/ marqueur modal + 0/ marqueur de personne + 0/interrogatif + 0/marqueur modal + 0/interrogatif + 0/ marqueur modal] *ou* [participe + 0/auxiliaire de temps + 0/temps + 0/ marqueur de personne + 0/interrogatif + 0/ marqueur modal + 0/marqueur de personne + 0/interrogatif] *ou* [participe + 0/déverbatif + fin nominal] *ou* [connectif + 0/interrogatif + 0/auxiliaire + 0/temps + 0/ marqueur de personne + 0/interrogatif] *ou* [déverbatif + fin nominale] *ou* [adverbiatif + 0/auxiliaire + temps + 0/marqueur modal + 0/marqueur de personne + 0/interrogatif] .

5.8 Règles de suffixation et cas spéciaux

Dans le cadre de cette étude préliminaire, nous nous sommes concentrés sur les formes verbales, car elles incluent aussi les suffixes utilisés pour les noms communs, comme on l'a vu dans l'analyse de structure ci-dessus. L'observation empirique nous a permis de définir les règles suivantes:

1. Seules les racines verbales monosyllabiques se terminant en *l* et contenant un *a*, *e*, *i* postérieur ou *o* peuvent être sujettes à une élision de consonne dans la forme connective et ses dérivés (par exemple: : aptu <= éliptu, kep <= kélip, qip <= qilip, bop <= bolup). Cette élision se produit essentiellement dans la langue parlée mais devient de plus en plus commune dans la langue écrite.
2. Les verbes monosyllabiques comportant un *a* dans la racine subissent un affaiblissement vocalique (*a*→*é*) seulement lorsqu'ils sont directement suivis d'une voyelle euphonique et restent inchangés dans les autres cas (par exemple: al- → él-i-p, él-i-sh, él-i-n-, mais al-i-men, al-i-dighan, al-ar). Par contre, l'affaiblissement vocalique se produit dès qu'un verbe monosyllabique comportant un *e* dans la racine (*e*→*é*) est directement suivi

¹⁶⁷ Les racines verbales peuvent être sujettes à des adoucissements ou déformations par exemple: kel- → kél- / ke-.

d'une voyelle, quelle qu'en soit la nature (par exemple: kel- → kél-i-p, kél-er, kél-i-men, kél-i-dighan, kéle)

3. La voyelle *é* affaiblie transmet la valeur antérieure ou postérieure de la voyelle originelle de la racine (par exemple: kél-i-shke, él-i-shqa).

4. Devant n'importe quel groupe CV, le *a* ou le *e* de la dernière syllabe d'un verbe polysyllabique s'affaiblit en un *i* qui transmet l'harmonie vocalique. (par exemple: sözle + mek → sözlimek; qatnash+ishKA → qatnishishqa) .

5. Les verbes polysyllabiques comportant une voyelle arrondie ou un *i* dans la dernière syllabe de leur racine se transforment comme les verbes monosyllabiques comportant une voyelle arrondie dans leur racine puisque ni les voyelles arrondies ni *i* sont affectés par l'affaiblissement vocalique.

6. Les verbes se terminant en une voyelle suivie d'un -n se comportent comme les verbes dans la forme réfléchie, dans le sens où ils n'assument ni la forme réflexive ni la forme passive. Le verbe *min-* (monter [à cheval, à bord]) semble être l'unique verbe de ce type qui puisse être utilisé à la forme passive.

7. Les verbes se terminant avec une double consonne ne sont pas affectés par l'affaiblissement, ex *yirt-* (avec un *i* postérieur).

8. Les deux verbes à monosyllabiques *de-* et *ye-* ajoutent un -y- euphonique devant les suffixes -sh- (nom verbal et pluriel) qui doivent alors être précédés d'une voyelle euphonique -i-.

9. Le verbe monosyllabique *yu-* ajoute un -y- euphonique devant les suffixes -n, -l, -sh, -sh-, -p, et -w, eux-mêmes précédés de la voyelle euphonique -u- (ex. yuyun-, yuyul-, yuyush, yuyush-, yuyup, yuyuw-). Il ajoute aussi un -y- devant le marqueur du temps du présent -i- (ex. yuyidu, yuyidighan). Il est souvent retrouvé dans la forme infinitive *yuymaq*, bien que l'orthographe officielle soit *yumaq*.

5.9 Les expériences et les résultats

Afin d'analyser les fonctions morpho-syntactiques du ouïghour, nous avons effectué l'expérience décrite ci-dessous. Notre approche prend en compte l'information de la structure du langage et les opérations de traitement de langue. Pour commencer, nous avons compilé une liste de 9265 verbes sur la base de laquelle nous avons établi un dictionnaire de

racines qui répertorie des entrées uniques non-suffixées, ou racines de verbes primitives. Puis nous avons créé une banque de règles¹⁶⁸ qui répertorie tous les suffixes verbaux dans leur ordre naturel d'apparence. Les suffixes ont été isolés basés sur l'approche de deux niveaux¹⁶⁹, appliqués aux caractéristiques spécifiques de l'ouïghour, et les marqueurs « fonctionnels » (marqueurs syntactiques) différenciés des lettres euphoniques. Par la suite, nous avons mis en place un générateur lexical¹⁷⁰ qui donne toutes les formes de surface théoriquement correctes de chaque verbe. Finalement, nous avons développé un explorateur de suffixes¹⁷¹ afin de visualiser/afficher la structure de chaque forme suffixée d'un verbe. La section suivante décrit chaque module en détail.

5.9.1 Dictionnaire des racines

Comme mentionné ci-dessus, le dictionnaire des racines ne contient que des racines verbales primitives¹⁷². Le *parsing/parsage*¹⁷³ des verbes dérivés de racines nominales existantes (qui représente la majorité des verbes répertoriés dans la liste de verbes d'origine) ex. *güllenmek* (s'épanouir, prospérer), *ajizlashmaq* [s'affaiblir, échouer, diminuer] montre bien qu'il s'agit de formes suffixées (dénominatives) de noms communs (ex. *gül*, fleur) ou d'adjectifs (ex. *ajiz*, affaibli). Les verbes en forme factitive irrégulière (ex. *keltür-* ou *kelgüz-* au lieu de la forme attendue *keldür-*, ou *chiqar* au lieu de *chiqtur*) n'ont pas été inclus puisqu'ils seront étudiés plus tard dans notre recherche, dans le cadre des règles de dérivation. Il doit être pris en compte que ces deux catégories de formes de verbes, même si elles sont exclues du dictionnaire des racines, suivent le même schéma de suffixation que les racines verbales primitives dans leur conjugaison. Toute entrée dans le dictionnaire inclut la racine, le marquer l'antériorité ou de postériorité vocalique et /ou l'information sur le type de mot. Exemple :

¹⁶⁸ Microsoft Excel, pour des raisons pratiques.

¹⁶⁹ Kemal Oflazer, 1994, *Two-level Description of Turkish Morphology*, Literary and Linguistic Computing, Vol. 9, No:2.

¹⁷⁰ Ceci est mis en œuvre en VBA.

¹⁷¹ Cet outil est mis en œuvre en C#, utilisant la composante TreeGX de devcomponents.com. Pour des raisons de droit d'auteur il n'est pas encore disponible en ligne mais une version prototype sera disponible lors de la présentation.

¹⁷² A ce stade, nous avons identifié 639 racines verbales primitives.

¹⁷³ Voir chapitre 9 pour plus de détails.

béqin-	V
til-	V
ich+	V
kel	V
sat	V

Dans la plupart des cas, la valeur d' antériorité ou de postériorité vocalique peut être vérifiée à partir de la dernière voyelle de la racine. Pour certaines racines qui ne contiennent que l'une des voyelles non-arrondies *é* et *i* et aucune indication de la valeur d' antériorité ou de postériorité de la racine, nous avons rajouté un + pour l'antériorité et – pour la postériorité. La lettre V suivie d'un caractère tabulation indique que la racine appartient au groupe verbal. Dans une prochaine étape, les racines nominales seront ajoutées et identifiées par la lettre N suivie d' un caractère tabulation.

5.9.2 Banque de règles

La banque de règles interprète la structure morphologique de chaque mot et est donc le résultat d'analyse linguistique. Elle est basée sur une harmonie vocalique à quatre sens (cf. section 5 harmonie vocalique) et les règles morpho-phonologiques du ouïghour. Elle est totalement indépendante du générateur lexical. La banque de règles consiste en 236 colonnes et 24,294 lignes¹⁷⁴. Les trois premières colonnes indiquent les changements affectant la racine¹⁷⁵, tandis que les colonnes suivantes représentent un suffixe chacune. Tout suffixe syntaxique probable¹⁷⁶ a été inclus ainsi qu'un certain nombre de suffixes déverbaux communs¹⁷⁷.

¹⁷⁴ Chaque colonne correspond à un suffixe. Le nombre de lignes indique le nombre de nouveaux mots théoriquement possibles qui peuvent ou non être utilisés dans la langue parlée.

¹⁷⁵ Voir aussi figure 5.1. La racine peut changer selon le contexte vocalique. Ex. “*kel*” a deux autres formes –“*kél*” {*kélemdu*} et “*ke-*” {*keptu*}.

¹⁷⁶ Incluent des suffixes verbaux qui sont absents des tables de formes verbales traditionnelles, tel le marqueur poli du pluriel.

¹⁷⁷ D'autres suffixes verbaux seront ajoutés en liaison avec le travail sur la dérivé des noms communs.

5.9.3 Générateur lexical

Le générateur lexical utilise le dictionnaire des racines afin de produire des nouvelles formes verbales en associant les racines avec ses suffixes possibles (voir fig.5.2). La liste contenant ces suffixes a été créée basée sur la banque de règles en disséquant, colonne par colonne, les suites de suffixes correctes. La combinaison de suffixe correcte est obtenue en appliquant les règles de suffixation. Les cas spéciaux (cf. section 8) sont pris en compte.

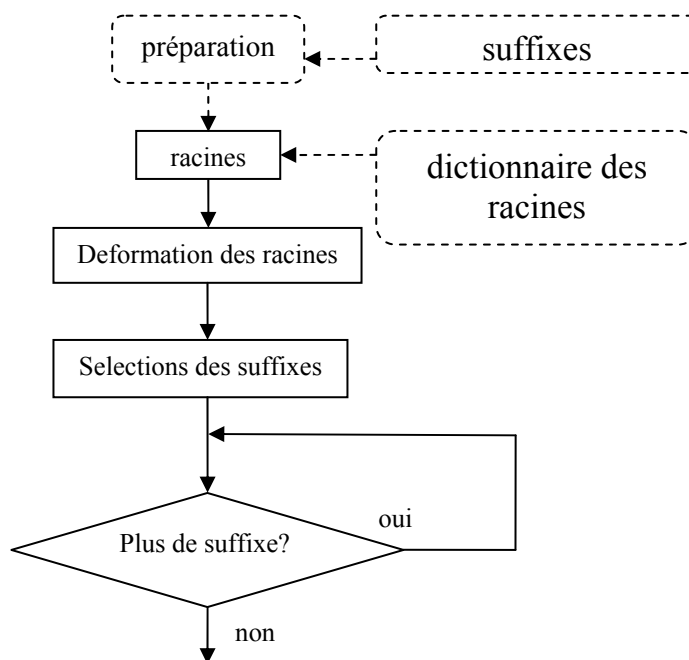


Figure 5.2. le générateur lexical

Afin d'obtenir les combinaisons de suffixes correctes lors de l'affixation, un mélange d'harmonies à deux sens et à quatre sens est appliqué, ainsi que des fonctions de douceur / dureté de consonne. Les suffixes transformateurs et de voyelles neutres sont aussi pris en compte et les règles euphoniques appliquées lorsque c'est nécessaire. Les exemples suivants démontrent des procédures de suffixation simples :

A. RACINE+ PARTICIPEPASSE +DENOM I*I*K:

{bar-}+{-ghan/qan/gen/ken}+{-liq/lik/luq/lük} + {din/tin}

(1) L'harmonie vocalique à deux sens élimine *gen* et *ken*; (2) l'harmonie des consonnes élimine *qan* parce que *q* est une consonne dure (cf. section 5.4.4.b); (3) l'harmonie à quatre

sens élimine *luq*, *lük* et *lik*; (cf. 5.4.6) ; (5) l'harmonie consonantique de D initiale élimine *tin* puisque la lettre finale du suffixe précédent est une consonne dure qui fait appel à une consonne dure (cf. 5.4.4.a). Finalement, le générateur obtient *barghanliqtin*, une forme correcte en ouïghour.

B. RACINE(C finale)+ CONNECT(C initial):

{bar}+I+{p}

(1) La consonne de fin de racine et le suffixe qui comporte une consonne initiale exigent un I euphonique. (2) Le I euphonique est un *i* postérieur d'après l'harmonie à quatre sens. (3) La présence d'une voyelle euphonique affaiblit la voyelle arrondie dans la racine. Finalement, le générateur obtient *bérip*, une forme correcte en ouïghour.

5.9.4 Explorateur de suffixes

Nous avons développé un explorateur de suffixes afin de présenter et d'analyser la structure de suffixation de l'ouïghour. Les suffixes sont classifiés en 60 catégories selon leurs caractères morpho-syntactiques (voir annexe 3). La forme de surface et frontière de suffixe de chacun est obtenu par la sortie du générateur lexical – des mots nouveaux et distincts. Dans le diagramme « arbre », les suffixes ayant un même rôle ne sont représentés qu'une fois. Etant donnée la complexité des schémas de suffixation, la structure entière est devenu un très grand image (1727 x 459381 pixel), qui représente 24,294 nouveaux mots dérivés de la racine « *kel* ».

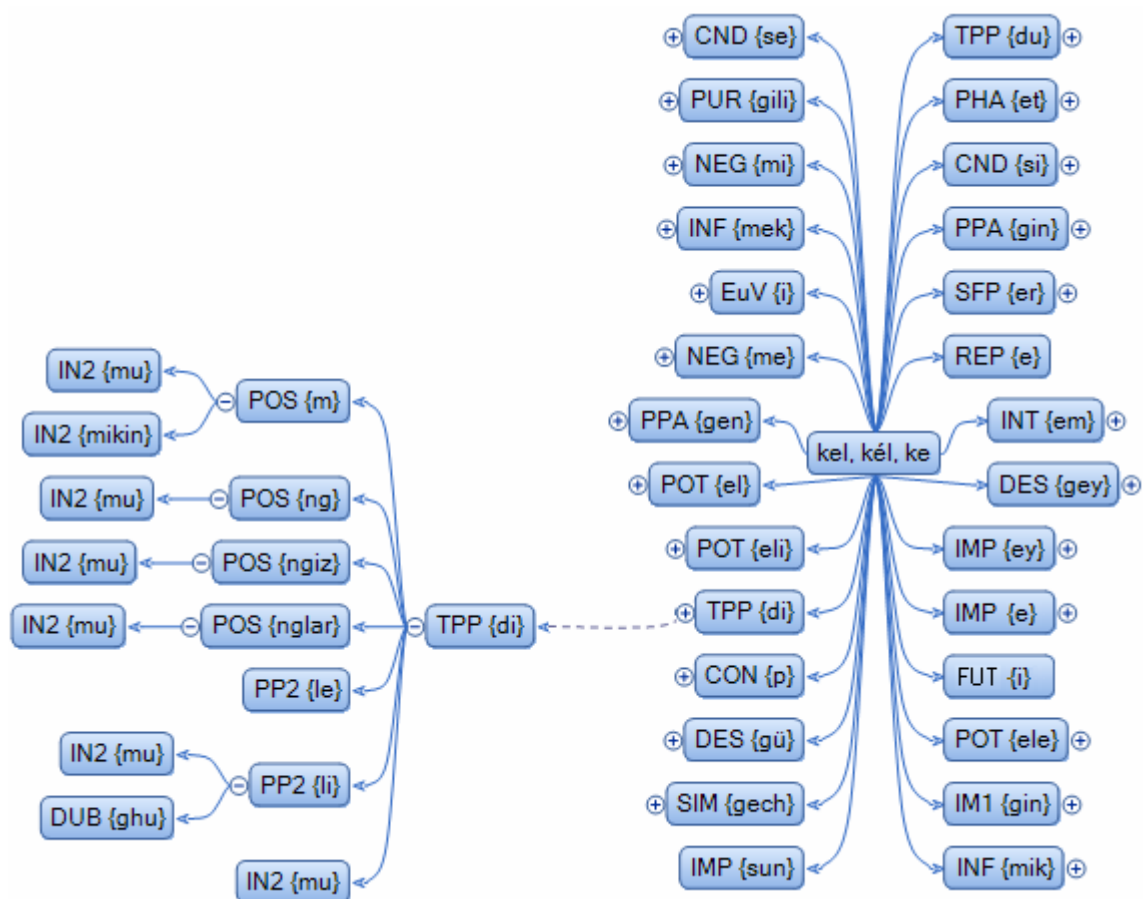


Figure 5.4, explorateur de suffixes, branches de «keldi » Figure 5.3. explorateur de suffixes niveau 1.

Le premier niveau de la structure de suffixation des mots ouïghours est démontré dans la figure 3. Un symbole « + » du côté gauche ou droit indique que plus de suffixes peuvent être ajoutés au morphème précédent. Nous avons déterminé que 13 est le nombre maximum de « rameaux » possible. La figure 4 démontre tous les rameaux commençant par la branche « *di* », à partir de laquelle nous pouvons obtenir les mots suivants : *keldi*, *keldim*, *keldimmu*, *keldimmikin*, *kelding*, *keldingmu*, *keldingiz*, *keldingizmu*, *keldinglar*, *keldinglarmu*, *keldile*, *keldilimu*, *keldighu*, *keldimu*.

Chapitre 6. Morphologie de l'ouïghour

Dans ce chapitre, nous présenterons la structure phonétique, syllabique et morphologique de la langue ouïghoure. Nous proposerons une méthode de sélection des suffixes afin de générer des noms conformes aux règles phonétiques de la langue. Par ailleurs, afin d'appliquer les notions de la structure morphologique de l'ouïghour au traitement informatique, nous analyserons les éléments agglutinants qui créent des noms à partir de la racine.

6.1 Compréhension et connaissance

Avant d'entrer dans la description détaillée, prenons un exemple pour analyser la morphologie de l'ouïghour. Prenons une suite de sons en ouïghour, que nous représenterons par :

Gyllyklirinqizdikilærdinmidi ? (*Translittération en Alphabet phonétique international*)

Nous comprenons qu'il s'agit d'un énoncé complet en ouïghour. Nous pouvons en donner une version écrite :

گۈللۈكليرىڭىزدىكىلەردىنمىدى؟

[Güllükliringizdikilerdinmidi ?] (*translittération en LSU, voir annexe 4*)

Pourquoi peut-on dire qu'on comprend cet énoncé ? Quel est le processus suivi pour ce faire ? Tout d'abord, il est nécessaire de connaître l'ouïghour. Cela implique que nous puissions mettre en œuvre tout un ensemble de connaissances liées à la compréhension d'une langue. En particulier :

— Nous savons associer à la suite de sons

*gyl+lyk+li+ri+qiz+di+ki+lær+din+mi+di*¹⁷⁸ une suite de racines et de suffixes

گۈل+لۈك+لەر+دىڭىز+دە+كى+لە+رىڭىز+دىن+مى+دى (gül+lük+ler+ingiz+de+ki+ler+din+mi+di). Pour cela, il faut être capable de découper et éventuellement regrouper la suite de sons en unités distinctes, qui sont ici /göl/, /lök/, /ler/, /ingiz/, /de/, /ki/, /ler/, /din/, /mi/ et /di/, et qui sont formées de ce que nous savons être des sons de l'ouïghour. Par exemple : [g], [y], [ŋ], [æ],

¹⁷⁸ Le signe « + » est ici utilisé pour séparer les unités lexicales les unes des autres.

[r], etc. Dans cet exemple, nous savons, en outre, interpréter le [i] des suffixes *lir*, *di* comme un phénomène d'affaiblissement de la voyelle *e* vers *i*. Les connaissances que nous mettons ici en œuvre sont des connaissances phonétiques et phonologiques.

— Il faut également relier ces unités à des mots de l'ouïghour. Cela suppose en premier lieu que nous sachions reconnaître dans chaque unité une forme d'un mot de référence, ou lemme, que l'on pourrait retrouver dans un dictionnaire. Ainsi, nous reconnaissons dans گۈللۈك (*güllük*) la racine گۈل (*gül*, fleur) à partir de laquelle on a créé le mot *güllük* (jardin) avec l'aide d'un suffixe لۈك (*lük*, lieu où quelque chose abonde). Nous savons aussi que le suffixe لەر (*ler*) est ajouté pour le pluriel —*güllükler* (jardins). Nous reviendrons plus en détail sur les explications des suffixes dans les paragraphes suivants. Si nous cherchons dans un dictionnaire l'entrée گۈل (*gül*), nous verrons qu'il s'agit d'un nom -- fleur. Nous trouverons dans la même page des mots composés ou des mots qui sont formés en rattachant des morphèmes. Par exemple : گۈللۈك (*güllük*, jardin), گۈلدەستە (*güldeste*, bouquet de fleurs). Mais, nous ne trouverons pas toutes les formes fléchies car elles peuvent être très nombreuses. Nous avons donc utilisé ici des connaissances morphologiques.

Arrivés à ce niveau de compréhension, nous avons les éléments principaux de la morphologie. Si nous continuons d'analyser les sens d'un mot dans différentes phrases, nous mettons en œuvre des connaissances sémantiques, qui ne font pas l'objet de notre recherche.

L'exemple ci-dessus illustre le fait que la compréhension d'une phrase, fût-elle très simple, met en jeu un ensemble de connaissance qui relèvent du domaine — connaissances linguistiques. Ce sont les connaissances dont nous venons de parler, qui sont directement liées à la langue elle-même, à son matériau phonique ou écrit, à sa structure, ainsi qu'à ses fonctions de référence à la réalité : connaissances phonétiques, phonologiques et morphologiques.

گۈل + لۈك + لەر + بگنر + دى + كى + لەر + دىن + مى + دى¹⁷⁹

¹⁷⁹ C'est la version ouïghour modifiée de l'exemple donné par Richard Sproat pour le turc. *Morphology and Computation*, page 44, The MIT Press, 1992.

Gül + lük + lir + ingiz + di + ki + ler + din + mi + di (de droite à gauche)
 (« Fleur » + DER + PL + 2SG.POS + LOC + REL + PL + ABL + INT + VT)
 « Sont-ce celles de vos jardins ? »

Dorénavant, comme la plupart de nos lecteurs ne sont pas ouïghourophones, afin de simplifier la lisibilité des mots ouïghour nous n'écrirons que la translittération standardisée en écriture latine – UKY. Voir l'annexe 4 pour l'équivalence des lettres ouïghour.

6.2 Structure phonétique

6.2.1 Les voyelles

La langue ouïghour comporte 8 voyelles. Nous les présentons ici sous forme d'un tableau qui indique leur position et la forme des lèvres correspondante.

Forme Position	pas arrondie			Arrondie	
	antérieure	médian	postérieure	antérieure	postérieure
Haute		ى i		ئۇ ü	ئو u
médiane		ې é		ئۆ ö	ئو o
Basse	ئە e		ئا a		

Tableau 4 : les voyelles

6.2.2 L'affaiblissement des voyelles

Quand on ajoute un suffixe à un mot monosyllabique, la première syllabe devient une syllabe ouverte et l'accent tonique se déplace pour tomber sur la deuxième syllabe. Si la voyelle du mot est un « a » ou un « e », elle s'affaiblit alors et devient un « é ». Par exemple :

at (n. cheval) + *i* (3p.POS) = *éti* (son cheval), *en* (largeur) + *i* = *éni* (sa largeur).

Quand on ajoute un suffixe à un mot polysyllabique, la dernière syllabe perd son accent tonique et si elle contient un « a » ou un « e » celui-ci s'affaiblit et devient un « i ». Par exemple :

ata (père) + *si*(3p. POS) = *atisi* (son/leur père), *yasa* (fabriquer, faire) + *ghan* (3p. passé composé) = *yasighan* (fabriqué, fait), *sözle* (parler, dire) + *gen* (TEMPS-P2) = *sözligen* (il a parlé, dit).

Ce phénomène ne se produit pas dans les mots d’origine arabe et persane où le « a » ou le « e » de la dernière syllabe est long (soit de nature , soit en raison de la présence d’un hamze ou d’un ‘aïn disparu de la graphie ouïghoure moderne). Par exemple : *imza* (a : إمضاء , signature) + *si* = *imzasi* (sa signature), *binakar* (p : بناكار, architecte) + *i* = *binakari* (son architecte).

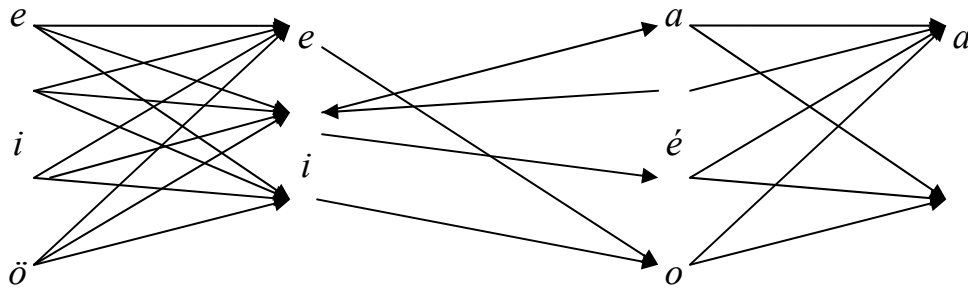
6.2.3 L’harmonie vocalique

L’utilisation des suffixes suit les règles de l’harmonie vocalique. [Reinhard F. Hahn & Ablahat Ibrahim, 1991] présentent celles-ci dans le cadre de la langue parlée, [Mirsultan Osmanov, 1985] le fait dans celui de la langue écrite. [A. Qaydarov & Gh. Sedwaqasov & T. Talipov, 1963] s’intéressent aux deux. Nous tentons ici de montrer les règles applicables au TAL en résumant le système vocalique de l’ouïghour :

1. La variation de la voyelle du suffixe se fait en fonction de la voyelle de la dernière syllabe du mot auquel il se lie. Dans plupart de cas, suivant les conventions qui gouvernent les rapports intervocaliques, le suffixe comporte une voyelle de même type (front, milieu, arrière) de voyelle que la dernière syllabe du mot¹⁸⁰.

v. de le dernière syllabe	voyelle du suffixe	v. de la dernière syllabe	voyelle	du
suffixe				

¹⁸⁰ L’ouïghour a absorbé un grand nombre de mots d’origine étrangère—surtout de l’arabe et du persan, mais aussi du mongol, du russe, du chinois et, plus récemment, d’autres langues. Phonologiquement, ces mots se comportent différemment des termes proprement türks de l’ouïghour, particulièrement en ce qui concerne les règles d’harmonie vocalique interne. Mais ils se plient aux règles d’harmonie dans leurs rapports avec les suffixes.



Par exemple : *geme* + *de* (dans la cave), *ter* + *i* (sa sueur), *sez* + *gü* (sensation), *köl* + *ge* (au lac), *ös* + *üm* (eco : intérêt, croissance), *idare* + *ler* (bureaux), *san* + *lar* (chiffres).

Il n'y pas de suffixe comportant *é* et *ö*. Il y a un seul suffixe qui comporte la voyelle *o*. Il s'agit du suffixe persan « *xor* », qui peut, comme dans sa langue d'origine, s'attacher à n'importe quelle syllabe sans transformation. Très peu de mots sont dérivé en utilisant « *xor* » et plupart ont des sens négatif. Par exemple : *gösh* (viande) + *xor* = *göshxor* (mangeur de viande), *para* (argent sale) + *xor* = *parixor* (corrompu), *qan* (sang) + *xor* = *qanxor* (vampire), *miras* (héritage) + *xor* = *mirasxor* (héritier ; coureur d'héritages).

2. Les mots monosyllabiques contenant « *i* » ainsi que les mots polysyllabiques dont la dernière syllabe contient un « *i* » prennent des suffixes comportant des voyelles antérieures – *e*, *ü*, *ö*—s'ils contiennent également « *g* » ou « *k* » (le suffixe relationnel –*ki* s'inscrit également dans ce cas de figure) ; sinon ils prennent des suffixes comportant des voyelles postérieures – *a*, *u*, *o*. par exemple :

Tik + *ken* = *tikken* (planté), *kigiz* + *ler* = *kigizler* (feutres), *til* + *lar* (les langues), *din* + *gha* = *dingha* (à la religion), *yiraq* + *ti* + *ki* + *ler* (ceux qui sont loin).

exceptions : *biz*, *siz*, *tiz*, *iz*, *chish*, *ich*, *jiddiy*, *ilmek* (crochet), *bil-*

3. Si la syllabe précédant un nouveau suffixe contient « *i* », ce suffixe conservera l'harmonie vocalique avec la dernière syllabe donnant une indication sur l'articulation vocalique. Par exemple :

tilliri (leurs langues) + *da* = *tillirida* (dans leurs langues), *yaghachchi* (celui qui travaille le bois—menuisier) + *gha* = *yaghachchigha* (au menuisier), *güli*(sa fleur) + *ge*(à) = *gülige*(à sa fleur), *güli* + *de* = *gülide* (dans sa fleur).

4. L'affaiblissement vocalique cause une ambiguïté, particulièrement dans les mots à racine monosyllabique, puisque une lettre *é* remplace aussi bien un *a* affaibli qu'un *e* affaibli¹⁸¹. Les suffixes, quant à eux, maintiennent l'harmonie vocalique comme si les voyelles n'avaient pas été affaiblies. Par exemple :

<i>bar</i> (V+0, aller)	+ <i>ish</i>	= <i>bérish</i> (NV : aller/donner)	+ <i>qa</i> / <i>ke</i>	bérishqa (pour/à aller)
<i>ber</i> (V+0, donner)				bérishke (pour/à donner)
<i>at</i> (V+0, tirer)	+ <i>ish</i>	= <i>étish</i> (NV : tirer/ faire)	+ <i>qa</i> / <i>ke</i>	étishqa (pour/à tirer)
<i>et</i> (V+0, faire)				étishke (pour/à faire)

5. Si la dernière syllabe est «*che*», un suffixe diminutif ou qui indique la manière et par extension la langue, la règle 1 s'applique à la syllabe qui précède le «*che*». Par exemple :

Kitab-che (livret) + *gha* = *kitabchigha*(au livret), *qelem-che*(bouture) + *ge* = *qelemchige*(à la bouture), *uyghur-che* (la langue ouïghoure) + *gha* = *uyghurchigha* (en ouïghour), *ereb-che* (la langue arabe) + *ge* = *erebchige* (en arabe).

6.2.4 Les consonnes

La langue ouïghoure comprend 24 consonnes. Nous les présentons ici sous forme d'un tableau¹⁸² en montrant leur position et leurs points d'articulation :

¹⁸¹ Ce problème d'ambiguïté au niveau de la graphie dans toutes les écritures ouïghoures ne reflète pas exactement la réalité phonétique de la langue parlée qui réalise différemment le «*é*» suivant qu'il a été formé à partir d'une voyelle antérieure ou postérieure. Il est à l'origine de beaucoup d'hésitations entre les graphies «*é*» et «*i*» et a même fait l'objet de quelques réformes orthographiques partielles dans les années 60 (cf Amine p. 24).

¹⁸² Ce tableau est un résumé des tableaux repérés de trois livres publiés par [Reinhard F. Hahn & A. Ibrahim, 1991, États-Unis, A Qaydarov & Gh. Sedwaqasov & T. Talipov, Almatï, 1963, R. Imin, 2000, Ürümqi].

articulation	position cordes vocales	labiale s	alvéo- dentales	alvéo- palatale s	Vélaire s	uvulaires	laryngales
occlusives	sonores	ب <i>b</i>	د <i>d</i>		گ <i>g</i>	ق <i>q</i>	
	sourdes	پ <i>p</i>	ت <i>t</i>		ك <i>k</i>		
affriquées	sonores		ج <i>j</i>				
	sourdes		چ <i>ch</i>				
spirantes	sonores	و <i>w</i>	ز <i>z</i>	ژ <i>zh</i> , ي <i>y</i>	غ <i>gh</i>		ه <i>h</i>
	sourdes	ف <i>f</i>	س <i>s</i>	ش <i>sh</i>	خ <i>x</i>		
nasales	sonores	م <i>m</i>	ن <i>n</i>		ڭ <i>ng</i>		
latérale	sonores		ل <i>l</i>				
vibrantes	sourdes		ر <i>r</i>				

Tableau 5. Les consonnes

Certaines de ces consonnes sont soumises à des règles d'harmonie ou d'assimilation phonétiques.

L'assimilation phonétique, qui n'est pas toujours représentée dans la graphie, affecte les consonnes sonores *b* et *d* qui deviennent respectivement les sourdes *p* et *t* en fin de mot. Elle affecte également en position finale les sourdes *k*, *q* et parfois *p* qui deviennent respectivement les sonores *g*, *gh* et *b* (ou même parfois *w*) lorsqu'on leur ajoute un suffixe commençant par un *i*.

6.3 Structure syllabique

6.3.1 Composition syllabique

La structure syllabique des mots ouïghours est très variée. Il y a des mots qui se composent d'une seule syllabe, et il y a des mots qui en contiennent plusieurs. En ouïghour contemporain, une syllabe peut suivre d'une des formes syllabiques suivantes :

Notes : C – consonne, V – voyelle. Les signes « + » sont utilisés pour séparer les différentes syllabes.

1. Une seule voyelle (V).

u (lui), **a** + **ta** (père), **qa** + **i** + **de** (régulation), **e** + **te** (demain)

Notons dans ce cas et le cas suivant que la graphie arabo-persane de l'ouïghour fait précéder toute voyelle initiale d'un glyphe particulier, un *hamze*, qui, dans le corps d'un mot indique un hiatus entre deux voyelles (comme c'est le cas dans *qaide* ci-dessus). Ce *hamze* initial, qui correspond à la « dent » précédant toute voyelle initiale dans la vieille écriture ouïghoure (voir annexe 4) est-il une sorte de consonne zéro? On peut se le demander dans le cas de mots composés comme *üinalghu* (magnétophone) où le *hamze* initial du « a » de *al* (prendre) est conservé et permet de segmenter le mot ainsi : *ün+al+ghu*.

2. Une voyelle et une consonne (VC).

At (cheval), **on** (dix), **öy** (maison), **al** + **ma** (pomme)

3. Une consonne suivie d'une voyelle (CV).

Bu (ce), **cho** + **ka** (baguette), **mé** + **we** (fruit), **pel** + **to** (manteau).

4. Consonne + voyelle + consonne (CVC).

Mek+tep (école), **lesh** + **ker** (soldat), **tam** + **gha** (tampon), **mis** + **kin** (triste)

5. Une voyelle suivie de 2 consonnes (VCC).

Eyt (v. dire), **ast** (sous), **üst** (sus), **erz** (pétition)

6. Une consonne et une voyelle suivies de 2 consonnes (CVCC).

Xelq (peuple), **dost** (ami\e), **ders** (leçon)

Les 6 types de structures ci-dessus sont les formes les plus courantes et régulières de la construction des syllabes ouïghoures. Il y a aussi 5 autres formes qui sont utilisées plutôt pour les mots d'emprunt récents. A savoir : CCV, CCVC, CCVCC, CVV, CVVC. Ils ne suivent pas le règlement général de segmentation des mots en syllabes. Par exemple :

pla+nér (r : планёр, planeur), **gram+ma+ti+ka** (r : грамматика, grammaire), stansiye (r : станция, gare), **tran + sport** (transport), **front** (front), jung + **xua**(ch : 中华zhōnghuá, Chine), **shöyüen** (ch : 学院 xuéyuàn, collège).

Notons en passant que, dans la langue parlée, ces mots sont transformés, pour se soumettre aux formes syllabiques et aux règles de prononciation traditionnelles, en pi+la+nér, gi+ram+ma+ti+ka, is-tan-si+ye, ti+ra+nis+port, pi+ront, jung+xa, shö+yen.

Cette notion de la syllabisation est très utile au TAL ouïghour, plus particulièrement, pour des logiciels de correction orthographique, ROC, outil de dictionnaire, traduction à l'aide d'ordinateur, analyseur de la morphologie, moteur de recherche, extraction d'information.

6.3.2 Règlement de segmentation syllabique

Les règles de segmentation syllabique nous aident pendant l'identification des unités morphologique des mots ouïghours. Les règles ci-dessous sont applicables seulement pour les mots ouïghours, excluant certains mots d'emprunt. A savoir :

1. S'il y a une consonne entre deux voyelles, la consonne s'associe à la deuxième voyelle pour composer une syllabe. Par exemple :

ata (père) = **a** + **ta**, **börek** (rognon) = **bö** + **rek**, **melike** (princesse) = **me** + **li** + **ke**

2. S'il y a deux consonnes entre deux voyelles, les deux consonnes se séparent en s'associant chacune à la voyelle la plus proche. Par exemple :

mektep(école) = **mek** + **tep**, **saylam**(élection) = **say** + **lam**, **oyman** (basin) = **oy** + **man**

3. S'il y a trois consonnes entre deux voyelles, les deux premières consonnes s'accolent à la première voyelle, la troisième consonne s'accole à la deuxième voyelle. Par exemple :

dostluq (amitié) = **dost** + **luq**, **xelqning**(du peuple) = **xelq** + **ning**

Toutes ces règles seront utilisées pour la dérivation des nouvelles entités lexicales à partir des racines et dans le respect de l'harmonie vocalique. Il est aussi important de factoriser les complexités introduites par la richesse de l'ouïghour en mots redoublés ou composés avec ou sans trait d'union qui demandent un traitement spécifique dans l'algorithme.

6.4 Structure morphologique

Les principes de la morphologie de l'ouïghour contemporain sont largement similaires aux autres langues turques. Par conséquent, ils sont aussi similaires aux autres langues agglutinantes de la famille altaïque par exemple : mongol, japonais, finnois¹⁸³. Le principe fondamental de la structure morphologique de l'ouïghour est de relier les séquences de suffixes (rarement de préfixes, puisque ceux-ci sont exclusivement empruntés au persan) à la racine d'un mot en créant des nouvelles formes lexicales. Ce principe est dit «agglutinant». Le résultat du processus d'agglutination dépend de l'apport syntactique de chacun des éléments qui y est incorporé.

6.4.1 L'aspect agglutinant des langues

« Une langue agglutinante est une langue dans laquelle les mots sont formés en rattachant des morphèmes ensemble. Proposée par le linguiste allemand Wilhelm Humboldt¹⁸⁴ en 1836 et reprise par son compatriote August Schleicher en 1861. Cette classification est encore largement utilisée pour classer les langues d'un point de vue morphologique. Les langues agglutinantes sont la forme la plus commune des langues synthétiques et elles sont habituellement hautement fléchies. Dans une langue agglutinante, les mots en général sont longs et complexes car ils se composent d'une série de morphèmes concaténés, comme un chapelet.

Le type agglutinant se caractérise par trois propriétés :

¹⁸³ Voir <http://sophistikatedkids.com/turkic/40%20Language/Lingo-Ethnical%20TreeEn.htm> pour la famille de la langue altaïque.

¹⁸⁴ http://www.wordiq.com/definition/Agglutinative_language

1. Une même unité lexicale peut avoir plusieurs formes selon les valeurs des éléments agglutinants qui lui sont associées. Ces variations de forme se font à l'aide de deux procédés :

- l'affixation (suffixation, préfixation, infixation)
- les alternances internes, vocaliques ou consonantiques.

2. Les signifiants correspondant aux différentes valeurs grammaticales ne sont pas autonomes les uns par rapport aux autres, ils sont fusionnés (ou amalgamés), voire même aussi fusionnés avec le signifiant de l'unité lexicale.

3. Les segments représentant les valeurs grammaticales correspondent à plusieurs valeurs distincts. [Jacques Poitou 2009 : site : <http://j.poitou.free.fr/pro/html/gen/typologie.html>] . Exemple :

- **ouïghour** : öy = maison, öy-ning = maison + génitif, öy-ler = maison + pluriel, öy-lir-im = maison + pluriel + possessif (1^{ère} personne : mon, ma, mes), öy-lir-im-ning = maison + pluriel + possessif + génitif

L'ouïghour est une langue agglutinante. Exemples de langues agglutinantes : finnois, hongrois, inuktituk, japonais, coréen, grec, toutes les langues turques (ex : azéri, kazakh, kirghiz, ouzbek, tatar, turc et turkmène), allemand, néerlandais, espéranto. [Pour de plus amples information, voir le site <http://j.poitou.free.fr/pro/html/gen/typologie.html>]

L'objectif de notre étude n'est pas d'analyser des langues par famille, nous allons présenter dans ce paragraphe les caractéristiques et la structure de la langue ouïghoure.

6.4.2 Type des éléments agglutinants

En ouïghour, comme en toutes les langues turques, la création de nouveaux mots de mots composés ou la composition de phrase entraînent de nombreux changements dans les mots. La liste des éléments agglutinants, plus particulièrement des morphèmes qui

s'attachent aux racines, se compose de deux catégories principales qui se distinguent selon leurs fonctions.

— **Morphèmes dérivationnels** : ils créent de nouvelles unités lexicales sur la base des unités préexistantes. Cette procédure de dérivation lexicale peut grandement augmenter le lexique de la langue. Il y a plusieurs méthodes pour produire des nouveaux mots. Par exemple :

1. *Gül* (n. fleur)+*lük* = *güllük* (n. jardin ; adj. fleuri) ; *kün* (n. soleil, jour) + *lük* = *künlük* (n. parapluie ; adj. Par jour). Ici, on a créé 2 nouveaux mots en ajoutant un suffixe *lük* — «qui a/ont». Chacun des mots a un sens comme un nom, et un autre sens comme un adjectif selon le contexte. A savoir :

Bu bir *güllük* (c'est un jardin). Bu bir *güllük köynek* (c'est une chemise à fleurs).

Bu bir *künlük* (c'est un parapluie). Bu bir *künlük tamaq* (c'est la nourriture pour un jour).

2. *élish* (prendre) + *bérish* (donner) = *élish-bérish* (relations); *qara* (noir) + *yüz* (face) = *qara yüz* (adj. Et n. perfide). Ici, on a généré de nouveaux mots en mettant deux mots ensemble. On remarque l'existence du trait d'union en ouïghour.

3. *Téz* (adj. Rapide) + *téz* (adj. Rapide)= *téz-téz* (adv. Rapidement). On peut créer aussi des nouveaux mots en répétant un mot et en insérant un trait d'union au milieu.

4. *Nan* (n. pain) – *pan* (Ø) = *nan-pan* (n. qqch comme du pain) ; *chay* (n. thé) + *pay* (Ø) = *chay-pay* (n. qqch comme du thé).

Nan-pan yigüng barmu ? (veux-tu manger quelque chose *comme du pain* ?)

Chay-pay ichküng barmu ? (veux-tu boire *du thé ou une autre boisson de ce genre* ?)

On peut reproduire également des nouveaux mots en rajoutant un mot qui n'a pas de sens mais qui a presque le même son que le mot auquel il est postposé avec un trait d'union. La partie après le trait d'union est une constante. Dès qu'on voit la première partie suivie du trait d'union, on peut trouver tout de suite la deuxième partie. Ainsi, après «nan-» on ne peut trouver que «pan». Ce genre de mots, dans plupart des cas, a le sens de «quelque chose

comme xxx » ou « quelque chose du même groupe que xxx ». Dans les deux exemples précédents, «*Nan-pan yégüing barmu ?*» ne veut pas forcément dire «veux-tu manger quelque chose comme du pain ? », mais plutôt « puis-je vous servir quelque chose » puisque, dans la tradition ouïghoure, on sert toujours le pain accompagné de thé.

Parmi les quatre méthodes de production des morphèmes dérivés, l'utilisation du suffixe est la plus fréquente.

- **Morphèmes syntactiques** : ils expliquent les relations syntactiques et des fonctions variées d'un mot dans une phrase. Ils ne participent pas à la création lexicale, ils ne changent pas le sens principal d'un mot, ils ajoutent des sens grammaticaux. En ouïghour, les suffixes indiquant le nombre, les cas, la possession et l'interrogation appartiennent à cette catégorie. Dans les formes nominales, l'ordre de base des morphèmes syntactiques est : le nombre (singulier ou pluriel) + le possessif + les cas. Analysons l'exemple du paragraphe 4.2.1.

گۈل + لۈك + لىر + ئىگىز + دى + كى + لەر + دىن + مى + دى

Gül + lük + lir + ingiz + di + ki + ler + din + mi + di (de droite à gauche)

(Fleur + SUFF + PL + 2P. POS + LOC + REL + PL + ABL + INT + IMP)

« Sont-ce celles de vos jardins ? »

1. Gül – racine
2. lük – suffixe dérivatif
3. lir – pluriel. Forme affaiblie de – *ler*.
4. ingiz – deuxième personne possessif (formel).
5. di – locatif. Forme affaiblie de – *de*.
6. ki – relatif.
7. ler – pluriel.
8. din – ablatif.
9. mi – interrogatif. Forme affaiblie de – *mu*.
10. di – imparfait

Ici, on remarque qu'un mot ouïghour peut être très long en raison de sa structure complexe. La langue ouïghoure repose sur un système de suffixes qui s'ajoutent aux noms,

pronoms, adjectifs, adverbes et verbes qui permettent d'exprimer un grand nombre de notions en peu de mots. C'est pour cela qu'un mot ouïghour devient plusieurs mots quand on traduit vers une autre langue.

6.4.3 Composition des mots ouïghours

Les mots ouïghours sont formés par les éléments constitutifs. L'élément qui se trouve en tête du mot est forcément la racine. Il s'ensuit que la forme traditionnelle d'un mot ouïghour est :

RACINE + UN ou PLUSIEURS SUFFIXES

Il n'existe que trois préfixes « *na-*, *bi-*, *bet-* ». Ils sont d'origine persane et sont employés, dans la langue recherchée, pour des mots d'origine arabo-persane ou, plus rarement, pour des mots empruntés à d'autres langues. Ils signifient « sans, in- » et sont remplaçables dans la langue courante par le suffixe « *-siz* ». Par exemple :

Na + *ilaj* (solution) = *nailaj* (insoluble) => *ilajsiz* ; *na* + *ümid* (espoir) = *naümid* (sans espoir, désespéré) => *ümidsiz* ; *bi* + *hajet* (nécessaire) = *bihajet* (inutile) => *hejetsiz* ; *bi* + *normal* = *binormal* (pas normal, anormal) => *normalsiz*.

Selon le lexique ouïghour, les mots sont constitués des quatre grands types suivants :

- **Racine** (ou stem) : Ce type de mots n'a qu'un seul morphème. Par exemple : *tagh* (montagne), *yer* (terre), *su* (eau), *tarix* (histoire), *namaz* (prière), *dijurné* (de service).
- **mots dérivé** : ce sont des mots qui sont formés en rajoutant des suffixes de dérivation aux stems. Par exemples : *ish* (travail) + *chi* = *ishchi* (travailleur), *küch* (la force) + *lük* = *küchlük* (fort). Dans certains mots empruntés d'origine iranienne, de nouveaux mots sont générés à l'aide de préfixes (par exemple : *bi* + *chare* [solution] = *bichare* [misérable], *na* + *ümid* [espoir] = *naümid* [sans espoir]) et de suffixes tels que *-xana*, *-xor*, *-shunas* et *-che*, aussi utilisés en ouïghour sur des racines d'origine non persane.
- **mots composés** : ce sont tout d'abord des mots composés de deux racines sémantiques accolées. Il ne s'agit plus de deux mots mais d'un mot composé

fixe, comme l'indique la graphie : apposition sans signe de subordination comme dans *tam* (le mur) + *gézit*(journal) = *tam gézit* (journal mural) ; trait d'union comme dans *mal* (bien) + *mülük* (possessions) + *mal-mülük* (patrimoine), ou incorporation en un seul mot, comme dans *ash* (nourriture) + *qazan* (casserole) = *ashqazan* (estomac) ou *töge* (chameau) + *qush* (oiseau) = *tögiqush* (autruche).

- **mots doublés** : Ce sont des mots formés de deux portions partageant certaines relations lexicales ou phonétiques et liées à l'aide d'un trait d'union. Par exemple : *qish* (hiver) + *yaz* (été) = *qish-yaz* (toute l'année) [2 mots du même domaine lexical], *bas-bas* (foule) [répétition du même mot], *chay* (thé) + *pay* (Ø) = *chay-pay* (boisson du type thé) [2 mots qui riment mais dont l'un n'a pas de sens individuel], *opur*(Ø) + *topur*(Ø) = *opur-topur* (précipitemment et en désordre) [deux mots qui n'ont pas de sens individuel].
- **abréviations** : Il y a deux manières de produire des abréviations en ouïghour : 1) écrire la première lettre de chaque mot suivie d'un espace. 2) coller ensemble la première syllabe de chaque mot (à la manière des communistes soviétiques). Par exemple :

Ouïghour	Ouïghour (UKY)	Française
ب د ت - بىرلەشكەن دۆلەتلەر تەشكىلاتى	BDT – Birleshken Döletler Teshkilati	ONU-Organisation des Nations Unies
ش ئۇ ئا ر - شىنجاڭ ئۇيغۇر ئاپتونۇم رايۇنى	ShUAR – Shingjang Uyghur Aptonum Rayuni	RAOX-Région Autonome Ouïghoure du Xinjiang
رايكوم - رايۇنلۇق كومىتېت	Raykom – Rayunluq Komitét	Comité Régional

Tableau 6. L'abréviation

Pour les noms de personnes, on peut abréger le prénom, voire le second prénom, en en gardant la première lettre que l'on fait alors suivre d'un point. Par

exemple, pour les deux noms de personnes – *Waris Abdukerim Janbaz* et *Attila Oghuz*, les abréviations ci-dessous peuvent être utilisées :

Waris A. Janbaz, W. A. Janbaz; A. Oghuz.

6.5 Constitution des noms

Le nom est un élément de base de la morphologie ouïghoure. Nous présentons les règles de la formation des noms en analysant les principes de deux grands éléments agglutinants – suffixes dérivatifs et suffixes syntactiques.

6.5.1 Suffixes dérivatifs de noms

Le suffixe est un instrument de formation des mots. La construction des noms à l'aide des suffixes dérivatifs est une méthode efficace largement utilisée. Dans ce paragraphe, nous présentons, en forme de tableau, les suffixes les plus couramment utilisés en donnant des exemples et la signification des mots créés. (Voir annexe 6 pour la liste complète des suffixes).

1. suffixes de noms personnels

Suffixes	liés aux	signification	Exemples
<i>-chi</i>	noms	agent, professionnel	<i>Gülchi</i> (fleuriste), <i>usul</i> (danse) + <i>chi</i> = <i>usulchi</i> (danseur), <i>süt</i> (lait) + <i>chi</i> = <i>sütchi</i> (laitier)
<i>-dash</i>	noms	celui qui partage la même chose	<i>yataq</i> (dortoir) + <i>dash</i> = <i>yataqdash</i> (colocataire), <i>isim</i> (nom) + <i>dash</i> = <i>isimdash</i> (qui a le même nom, homonyme)
<i>-ghuchi, -quchi, -güchi, -küchi</i>	verbes	agent	<i>Sat</i> (v : vendre) + <i>quchi</i> = <i>satquchi</i> (vendeur), <i>küt</i> (v : servir) + <i>küchi</i> = <i>kütküchi</i> (serveur)
<i>-liq, -lik, -luq, -lük</i>	noms propres	nationalité, origine	<i>Fransiye</i> + <i>lik</i> = <i>Fransiyelik</i> (français), <i>Lyon</i> + <i>luq</i> = <i>Lyonluq</i> (lyonnais)

Tableau 7. Dérivation des noms personnels

2. suffixes de choses et de noms concrets

Suffixes	liés aux	signification	Exemples
<i>-q, -k, -aq, -ek, -iq, -ik, -uq</i>	verbes	objet, endroit	<i>pich</i> (v : coupe) + <i>aq</i> = <i>pichaq</i> (couteau), <i>yat</i> (v : couche) + <i>aq</i> = (dortoir)
<i>-qa, -ke, -gha, -ge, -qu, -ghu</i>	verbes	objet	<i>chal</i> (v : joue) + <i>ghu</i> = <i>chalghu</i> (instrument de musique), <i>as</i> (v : crocher) + <i>qu</i> = <i>asqu</i> (crochet)
<i>-ghuch, -güch, -quch, -küch</i>	verbes	objet	<i>bas</i> (v : imprime) + <i>quch</i> = <i>basquch</i> (imprimante), <i>öchür</i> (v : efface) + <i>güch</i> = <i>öchürgüch</i> (gomme)
<i>-zar, shen</i>	noms	place	<i>Gül</i> + <i>zar</i> = <i>gölzar</i> (jardin), <i>üzüm</i> (raisin) + <i>zar</i> = <i>üzümzar</i> (vignoble), <i>göl</i> + <i>shen</i> = <i>gölshen</i> (jardin)

Tableau 8. Dérivation des noms concrets

3. suffixes de noms abstraits

Suffixes	liés aux	signification	Exemples
<i>-liq, -lik, -luq, -lük, -chiliq, -chilik</i>	adjectifs	état	<i>igiz</i> (haut) + <i>lik</i> = <i>igizlik</i> (hauteur), <i>yash</i> (jeune) + <i>liq</i> = <i>yashliq</i> jeunesse, <i>qiyin</i> (difficile) + <i>chiliq</i> = <i>qiyinchiliq</i> (difficulté)
<i>-liq, -lik, -luq, -lük</i>	noms	qualité	<i>küch</i> (force) + <i>lük</i> = <i>küchlük</i> (fort), <i>chöl</i> (désert) + <i>lük</i> = <i>chöllük</i> (désertique)
<i>-sh, -ish, -ush, -üsh</i>	verbes	nom d'action	<i>oqu</i> (n : lire) + <i>sh</i> = <i>oqush</i> (lire, lecture)

Tableau 9. Dérivation des noms abstraits

4. suffixe de minimisations, formes affectives et respectueuses

Suffixes	liés aux	signification	Exemples
<i>-chaq, -chek, -chuq, -chük, -chiq, -che, -jan, -laq</i>	noms, noms propres	minimalisation d'état	<i>Tay</i> (cheval) + <i>chaq</i> = <i>taychaq</i> (poulain), <i>yangaq</i> (noix) + <i>che</i> = <i>yangaqche</i> (noisette), <i>ana</i> (mère) + <i>jan</i> ¹⁸⁵ = <i>anajan</i> (ma petite maman), <i>bota</i> (chameau) + <i>laq</i> = <i>botilaq</i> (petit chameau)
<i>-ay</i> ¹⁸⁶ , <i>-gül</i> ¹⁸⁷ , <i>-m, -im, -um, -üm</i>	noms propres	affectif	<i>Tursun</i> (nom propre masculin) + <i>ay</i> = <i>Tursunay</i> (nom propre féminin), <i>Rizwangül</i> (nom propre féminin), <i>Waris</i> (nom propre masculin) + <i>im</i> = <i>Warisim</i> (nom propre masculin, « mon Waris »)
<i>-xan</i> ¹⁸⁸ , <i>-axun</i> ¹⁸⁹	noms propres	respectueuse	<i>Anar</i> (nom propre féminin) + <i>xan</i> = <i>Anarxan</i> (nom propre féminin, « respectueux »), <i>Nur</i> (nom propre masculin) + <i>axun</i> = <i>Nuraxun</i> (nom propre masculin, « respectueux»)

Tableau 10. Minimisations, formes affectives et respectueuses

5. suffixes persans

Les suffixes d'origine persane se joignent à des mots ouïghours, et aussi à des mots empruntés de l'arabe et du persan ou d'autres langues. Ils créent des noms concrets. En raison de l'irrégularité de l'utilisation des voyelles de suffixes persans, les règles de l'Arbre Janbaz ne peuvent s'appliquer sur le côté gauche.

¹⁸⁵ “-jan” est une forme de minimalisation respectueuse et amicale. Il se joint aussi aux noms propres un peu comme « Pierrot » pour le nom Pierre en français.

¹⁸⁶ « ay » - la lune.

¹⁸⁷ «gül » - fleur.

¹⁸⁸ “-xan” est une application directe du mot « xan (roi) comme un suffixe. Dans certains cas, on appelle des étrangères « Xan », par respect, si on ne connaît pas le prénom.

¹⁸⁹ “axun” – grade de théologien musulman bien éduqué (iranien Axund) ; utilisé comme suffixe, il indique le respect.

Suffixes	liés aux	signification	Exemples
<i>-kar, -ker, -gar, -ger</i>	noms, noms propres	agent	<i>bina</i> (architecture) + <i>kar</i> = <i>binakar</i> (architecte), <i>soda</i> (commerce) + <i>ger</i> = <i>sodiger</i> (commerçant)
<i>-kesh</i>	noms	agent	<i>pochta</i> (poste) + <i>kesh</i> = <i>pochtakesh</i> (vaguemestre), <i>kira</i> (loyer) + <i>kesh</i> = <i>kirakesh</i> (loueur)
<i>-dar,</i>	noms	agent	<i>emel</i> (position) + <i>dar</i> = <i>emeldar</i> (fonctionnaire)
<i>-xor</i>	noms	celui qui consomme	<i>para</i> (arjet) + <i>xor</i> = <i>parixor</i> (corrompu), <i>qan</i> (sang) + <i>xor</i> = <i>qanxor</i> (tueur), <i>ghem</i> (soucis) + <i>xor</i> = <i>ghemxor</i> (quelqu'un qui fait de soucis)
<i>-men</i>	noms d'action	celui qui est impliqué	<i>zoq</i> (profit) + <i>men</i> = <i>zoqmen</i> (amateur), <i>köch</i> (immigrer) + <i>men</i> = <i>köchmen</i> (immigrant)
<i>-shunas</i>	noms	spécialiste	<i>Til</i> (langue) + <i>shunas</i> = <i>tilshunas</i> (linguiste), <i>tarix</i> (histoire) + <i>shunash</i> = <i>tarixshunas</i> (historien)
<i>-pez, -wen</i>	noms	agent	<i>ash</i> (nuriture) + <i>pez</i> = <i>ashpez</i> (cuisinier), <i>derwaza</i> + <i>wen</i> = <i>derwaziwen</i> (gardien de porte)
<i>-xana</i>	noms	Pièce, endroit	<i>ash</i> + <i>xana</i> = <i>ashxana</i> (cuisine), <i>chay</i> + <i>xana</i> = <i>chayxana</i> (maison du thé)
<i>-dan</i>	noms	Ce qui contient	<i>sham</i> (bougie) + <i>dan</i> = <i>shamdan</i> (bougeoir), <i>kül</i> (cendre) + <i>dan</i> = <i>küldan</i> (cendrier)

-istan	noms	endroit	Özbékistan, Turkistan, qewristan (cimetière), gülistan (jardin fleuri)
--------	------	---------	---

Tableau 11. Suffixes iraniens

6.5.2 Suffixes syntactiques de noms

Nous présentons les suffixes syntactiques par trois aspects différents : le nombre, la possession et les cas.

6.5.2.1 Le nombre

En ouïghour, les suffixes « -lar » ou « -ler » expriment le pluriel. Les formes plurielles sont généralement formées à l'aide de ces deux suffixes. Par exemple :

Singulier	Pluriel
<i>sheher</i> (ville)	<i>sheherler</i> (villes)
<i>kitab</i> (livre)	<i>kitablar</i> (livres)
<i>er</i> (homme)	<i>erler</i> (hommes)
<i>ayal</i> (femme)	<i>ayallar</i> (femmes)

Tableau 12. le nombre

Dans certains cas, la singularité ou la pluralité ne sont pas exprimées et seul le contexte peut aider à faire disparaître cette ambiguïté. Par exemple :

Mende kitab bar. (J'ai un/des livre/s.) *Uning chishi yoq.* (il/elle n'a pas de dent/s.)

Afin d'éviter des ambiguïtés, on peut ajouter « bir » pour le singulier et indiquer le nombre exact de noms dénombrables pour le pluriel, mais les noms ne prennent alors pas la marque du pluriel. Par exemple :

Men on at we bir töge aldim. (J'ai acheté dix chevaux et un chameau.)

Et non *Men on at**lar** we bir töge aldim* qui serait une faute grave.

Dans une énumération, il suffit d'ajouter « -lar, -ler » seulement au dernier nom afin d'exprimer le pluriel. Par exemple :

*Mende alma, badam we anar**lar** bar.* (J'ai des pommes, des amandes et des grenades.)

(mais **pas** «*Mende almalar, badamlar we anarlar bar.*»)

Les deux suffixe « *-lar, -ler* » ajouter aussi d'autres sens que pluriel quand on les ajoute à différents noms. A savoir :

1. noms propres

Adillar keldi. (Adil et autres sont venus) *Béyjinglerge barduq.* (Nous sommes allés à Pékin et quelques autres villes.)

2. nom non-dénombrables

qushlar – différents types d'oiseaux.

Güller – différents types de fleurs

Derexler – différents types d'arbres

6.5.2.2 La possession

Les suffixes de possession sont variés et ils s'accordent avec la dernière lettre de l'unité qui représente ce qui est possédé. L'arbre Janbaz peut sélectionner les suffixes correspondant aux mots. Nous les présentons sous forme de tableau avec des exemples.

Dernière lettre Personne		Voyelle		Consonne	
		singulier	pluriel	Singulier	pluriel
1		<i>-m</i>	<i>-miz</i>	<i>-im/-um/- üm</i>	<i>-imiz</i>
2	Informel	<i>-ng</i>	<i>-nglar</i>	<i>-ing/-ung/- üng</i>	<i>-inglar/-unglar/-ünglar</i>
	Formel	<i>- ngiz, liri¹⁹⁰</i>	<i>-ngizlar / -ngizler/- liri</i>	<i>-ingiz, -liri</i>	<i>-ingizlar/-ingizler/-liri</i>
3		<i>-si</i>	<i>-si</i>	<i>-i</i>	<i>-i</i>

Tableau 13. la possession

¹⁹⁰ 2^{ème} personne formelle de la forme parlée formelle. Ex : *(özlirining) akiliri qandaqraq ? Ayallirichu ?* (Comment va votre frère ? et votre femme ?).

aka (frère) – akam, akimiz, akang, akanglar, akingiz, akingizlar, akisi
öy (maison) – öyüm, öyimiz, öyüng, öyünglar, öyingiz, öyingizler

Les suffixes possessifs nous aident à identifier le sujet dans les phrases où il n'est pas exprimé. Par exemple :

Soalim yoq (je n'ai pas de question) équivaut à *méning soalim* yoq.

6.5.2.3 Les cas

Les suffixes de cas indiquent la fonction des noms dans les phrases. Les noms subissent diverses modifications qui sont représentés par les suffixes de cas, selon le système de déclinaison ouïghour qui comprend les six cas suivants :

1. le cas zéro

Il ne présente aucun suffixe de cas. Il indique généralement le sujet de la phrase, par exemple :

Bu qiz keldi (**cette fille** est arrivée).

Il peut également indiquer un complément d'objet non défini. Par exemple :

Men kitab aldim (j'ai acheté UN livre/DES livres) par opposition à *Men kitabni aldim* (j'ai acheté le livre en question).

2. le génitif

Le signe du génitif est le suffixe « **-ning** ». Il indique la possession. Par exemple :

Türkistan ottura Asiyaning bir qismi (Le Turkéstan est une partie de l'Asie centrale).

3. le datif (directif)

Le suffixe « **-gha/ -qa/ -ge/ -ke** » est le signe du datif. Le datif indique la destination de l'action. Par exemple :

Men mektepke barmaymen, deslep sahilgha, andin öyge barimen, axirida chimliqqa chiqimen (Je ne vais pas **à** l'école, je vais d'abord **à** la plage, puis **à** la maison, enfin **sur** la pelouse).

Bu güllerni qiz dostumgha aldim (J'ai acheté ces fleurs pour ma copine).

4. l'accusatif

Il sert à désigner la personne ou la chose définie qui reçoit l'action marquée par le verbe. Il suffit d'ajouter « **-ni** » pour former ce cas. Par exemple :

Bu güllerni men ekelgen (J'avais amené ces fleurs).

5. le locatif

Le locatif est le cas qui localise l'endroit où l'action a lieu. Le suffixe « **-da/-de/-ta/-te** » est le signe du locatif. Par exemple :

Men Janbazda tughulghan (Je suis né à Janbaz).

6. l'ablatif

Il marque l'origine de l'action. Se mettent à l'ablatif les noms qui désignent le point de départ d'une action, le lieu par lequel on passe, d'où on s'éloigne et d'où on revient. Le suffixe de l'ablatif est « **-din/-tin** ». Par exemple :

Biz Qeshqerdin qaytip kelduq (nous sommes revenus de Qeshqer).

6.5.2.4 L'ordre des suffixes, de nombre, de possession et des cas

En ouïghour, le nombre des éléments agglutinants est infini. Leur ordre de base est PL+POS+cas+INT+TEMPS dans tous les cas, avec la possibilité d'absence de l'un ou l'autre des quatre types de suffixe et aussi l'effet que peut entraîner l'insertion du relationnel KI après un cas locatif ou génitif, qui permet d'ajouter un nouveau chapelet. Nous ne donnons qu'un aperçu des multiples possibilités.

Racine + {[Suffixes Dérivatifs]}[PL/SG][POS][CAS]} {REL} {[PL/SG][POS][CAS]}
 {[REL...]} {[INT][TEMPS]}

Alghiningiz mektep+ti+ki+ler+ning+ki+din+mu ? (Ce que vous avez pris est de ces qui sont dans notre école ?)

Mot + CAS+REL+PL+CAS+REL+CAS+INT

6.6 Radical

Le radical verbal peut prendre jusqu'à quatre formes :

1. Radical original : kel, sözle, al, kör, qoy etc.

2. Radical modifié par la présence d'une voyelle le suivant immédiatement (valable seulement pour les radicaux en e ou a qui se terminent pas une consonne—kel->kél+v ; al->él+v ; ber->bér+v) ou d'un groupe c+v (valable uniquement pour les radicaux qui se terminent un e ou a—sözle->sözli+cv ; qara->qari+cv) ou par la présence d'un groupe c-v le suivant immédiatement (valable uniquement pour les verbes terminés par une voyelle e ou a : sözle-> sözlidi ; qara->qaridi).

3. Radical étendu par l'ajout d'une voyelle euphonique (vE) devant certains suffixes commençant par une consonne, ce qui se combine pour les verbes en e/a avec la modification du radical (cf 2.) (valable seulement pour les radicaux se terminant par une consonne : kel->kéli+[sh/p/w], al->éli[sh,p/w], kör->körü[sh/p/w], qoy->qoyu[sh/p/w].

Les verbes monosyllabiques ouverts (sans consonne finale) ajoutent un y euphonique avant la voyelle euphonique ; la voyelle du radical, si elle est un e/a, se modifie (cf2) (de->déy ; ye->yéy ; yu->yuyu).

4. Radical modifié par l'élision du l final (valable uniquement pour les radicaux monosyllabiques se terminant en l : kélip-> kep ; élip->ap ; bolup->bop)

Les suffixes qui sont phonologiquement neutres (c'est-à-dire, qui ne transmettent pas la qualité avant ou arrière des mots) sont enregistrés une seule fois. Il s'agit de suffixes qui ne sont suivis que par des suffixes finaux invariables comme les pronoms personnels (men, sen, siz...) ou les marqueurs modaux (tiken, diken, mish). Il s'agit surtout du suffixe du passé simple de *ermek (imek), i.e. *di*, peut-être aussi de celui du présent (i,y)

Les suffixes phonologiquement actifs (c'est-à-dire, qui imposent une qualité vocalique aux suffixes suivants, comme *-dighan*, *-inglar...*) sont aussi enregistrés une seule fois.

Les suffixes phonologiquement transparents (c'est-à-dire, qui transmettent aux suffixes suivants la qualité vocalique des mots qui les précèdent) sont inscrits sous aux moins deux colonnes, une vf et une vb (ex : *-im* [+vf] (ex : *-im+-ge*) et *-im*[+vb] (ex : *-im+gha*)

Tous les suffixes sont inscrits sous les diverses formes qu'ils prennent en fonction de leur environnement phonétique (ex : *gen*, *gin-*, *ken*, *kin-*, *ghan*, *ghin-*, *qan*, *-qin-* ; *-m*[+vf], *-im*[+vf], *-üm*[+vf], *-m*[vb], *-im*[+vb], *-um*[+vb]).

L'ordre des suffixes dans le tableau doit suivre leur ordre d'insertion logique dans le mot. Le même suffixe peut apparaître plusieurs fois dû à la formation de verbes composés par l'utilisation de verbes auxiliaires incorporés au verbe. Ainsi, le suffixe –p/-w peut être utilisé pour la formation d'un verbe composé à la fin du radical et à la fin de l'auxiliaire (ex : kéliwétiwidim). Comme les suffixes marquant le temps (*i/y*, *se/sa*, *di/du/ti/tu* etc) ou la personne du verbe conjugué (*men*, *sen* etc.) n'apparaissent qu'une fois, ils sont regroupés vers l'arrière du tableau.

Le suffixe *di-/ti/du/tu/dur* précédant le pronom personnel final dans certains cas (ex : forme interrogative personnelle –*em/am+di+men*, forme honorifique –*di+le/la* etc.)) semblent être une survivance du chaghatay *dur* et ne sont pas un élément interne du pronom personnel final. Ils sont donc notés dans des colonnes séparées (*kél-+i+0+men* mais *kél-+i+du+0* ; *kél-+i+0+men* mais *kél-+em+di+men*).

Chapitre 7. Extraction de l'information ouïghoure

7.1 Les outils d'extraction

Ces dernières années, de nombreux logiciels sur le traitement automatique de la morphologie des langues naturelles ont été développés. Ils permettent de comprendre ou de constituer des composants linguistiques tels que les analyseurs morpho-syntaxiques et les systèmes de désambiguïsation lexicale sémantique. Nous souhaitons mentionner quelques outils existants :

INTEX¹⁹¹ : Laboratoire LASELDI, « *un analyseur de corpus basé sur l'utilisation de ressources lexicales à très large couverture. Il comprend plusieurs dictionnaires électroniques et des grammaires représentées par des graphes à états finis. Les utilisateurs peuvent ajouter leurs propres ressources au système. Ces outils sont utilisés sur des textes pour localiser des structures lexicales et syntaxiques, et également pour lever l'ambiguïté et l'étiqueter des mots simples ou composés.* » [Max SILBERZTEIN, « Outils pour le TAL », <http://www.atala.org/INTEX> , 2002]

TERMIUM®¹⁹² : une des plus grande base de données translinguistique (plus de 3,5 millions de termes), trilingue, du Bureau de la Traduction du Gouvernement Canadien, qui permet de repérer un terme en espagnol, en anglais ou en français à l'aide d'une requête dans l'une des trois langues.

FASTER¹⁹³ : Christian Jacquemin, un outil de reconnaissance de termes et de variantes pour l'anglais et le français. Il permet d'indexer des documents à partir d'un thésaurus ou d'une liste de termes récupérés d'une manière de l'indexation contrôlée.

LEXTER¹⁹⁴: Didier Bourigault, un analyseur syntaxique français dédié à l'extraction de syntagmes nominaux et adjectivaux à partir de corpus spécialisés, dans une perspective d'acquisition terminologique.

¹⁹¹ Max SILBERZTEIN, « Outils pour le TAL », Web <http://www.atala.org/INTEX> , 2002

¹⁹² <http://www.termium.com>

¹⁹³ <http://www.limsi.fr/Individu/jacquemi>

XFST, LEXC, TWLOC¹⁹⁵ : XRCE / Xerox, outils d'analyse morphologique multilingue permettent de faire une étude des suffixes de dérivation par transducteurs à nombre fini d'états. Disponible en 14 langues. La version turque est développée par Kemal Oflazer¹⁹⁶.

TTAT¹⁹⁷ : un outil d'annotation basé sur la morphologie du turc avec la fonctionnalité de désambiguïsation morpho-syntaxique. Il permet à l'annotateur/l'utilisateur de sélectionner manuellement la proposition adéquate s'il y a des ambiguïtés.

Unitex¹⁹⁸ : Sébastien Paumier - Université de Marne-la-Vallée, un ensemble de logiciels Open Source permettant de traiter des textes en langues naturelles en utilisant des ressources linguistiques. Il intègre des dictionnaires électroniques et différents composants nécessaires à l'analyse morphologique et syntaxique des textes du corpus.

7.2 Deux types d'extraction

Nous remarquons que les chercheurs ont deux grands types d'approches différentes pour le système d'extraction:

- « **Le point de vue statistique** concerne la distribution des termes dans les documents. Il se base sur l'hypothèse que l'emploi de la co-occurrence suggère une relation sémantique entre les termes. Ce point de vue a montré des limites dues essentiellement à l'ignorance du contexte linguistique » [Mohamed Hatem HADDAD, 2002]. [Oflazer & Tür, 2001] précisent que l'approche statistique donne un bon résultat pour le système d'extraction de l'information turc. Ils utilisent l'information lexicale et

¹⁹⁴ didier.bourigault@univ-tlse2.fr

¹⁹⁵ <http://www.xrce.xerox.com/competencies/content-analysis/fst>

¹⁹⁶ <http://www.nlp.cs.bilkent.edu.tr/cgi-bin/tmanew> (le site de référence n'est plus accessible.)

¹⁹⁷ *Turkish Treebank Annotation Tool*, Kemal Oflazer & Bilge Say & Dilek Zeynep & Gökhan Tür, <http://www.research.att.com/~gtur/pubs/ttbank-book.pdf>

¹⁹⁸ Logiciel gratuit sous licence GPL, disponible à l'adresse suivante : <http://www-igm.univ-mlv.fr/~unitex/download.html>

morphologique pendant la segmentation des phrases, topiques et entités nommées. Dans cette méthode, désambiguïsation s'effectue selon l'analyse statistique.

- « **Le point de vue linguistique** concerne les combinaisons des éléments textuels au niveau du discours. C'est un niveau proche de la syntaxe qui prend en considération les rapports syntagmatiques entre les éléments textuels. C'est une approche liée à plusieurs niveaux d'analyse linguistique – niveau lexicale, syntaxique et sémantique » [Mohamed Hatem HADDAD, 2002].

Le développement d'un outil complet demande énormément de travail. Ainsi, nous donnons trois algorithmes qui seront utiles pendant le traitement de l'ouïghour pour toutes les deux approches ci-dessus. Nous proposons notre choix technique pour mettre en œuvre ce modèle, particulièrement dans le cadre de l'extraction automatique des entités nommées.

7.3 Affichage des caractères

Les lettres ouïghoures ont été placées dans la région principale des lettres arabes (0600-06FF), dans la formule de la présentation arabe A (FB50–FDFF) et dans la formule de la présentation arabe B (FE70–FEFF) de la table de l'ISO 10646/Unicode.

Les codages actuellement utilisables sur les systèmes d'exploitations (Windows, Linux) pour l'ouïghour sont :

- ISO 8859-6 Latin/Arabic
- MS Arabic Code Page 1256
- Unicode
- Unicode (Big-Endian)
- Unicode (UTF-8)

L'exemple ci-dessous correspond aux codages des caractères ouïghours. Afin d'afficher correctement des caractères ouïghour, on peut inclure une des options suivantes dans la balise <head> d'un document HTML. Par exemple :

```
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-6">
```

```
<meta http-equiv="Content-Type" content="text/html; charset=windows-1256">  
<meta http-equiv="Content-Type" content="text/html; charset=unicode">  
<meta http-equiv="Content-Type" content="text/html; charset=unicodeFFFE">  
<meta http-equiv="Content-Type" content="text/html; charset=utf-8">
```

On peut également afficher un caractère ouïghour, indépendamment du codage de document HTML, en donnant la valeur qui correspond à la table Unicode.

A savoir : `&#xVALEUR UNICODE EN HEXADECIMAL;` ou `&#VALEUR UNICODE EN DECIMAL;`

Par exemple : `ب` ou `ب` correspond à la lettre ب.

Dans l'Unitex 1.0/1.1, il est impossible d'afficher 6 lettres Unicode (ؤ، و، ۇ، ڭ، ھ، ڭ، ڭ، ڭ، ڭ، ڭ) qui n'existe pas dans l'alphabet arabe. Ainsi, nous présentons l'extraction des entités nommées ouïghoures en utilisant l'alphabet latin.

7.4 Segmentation et traitement linguistique

La segmentation est un processus de découpage de texte en phrases, de phrases en séquence d'unités lexicales qui peuvent correspondre aux éléments de base morphologique. Dans les langues agglutinantes, les mots comportent plusieurs unités ensemble et ils sont en général complexes car ils se composent d'une série de morphèmes concaténés sans séparées par un espace, comme un chapelet. Traditionnellement, la segmentation prototype a été réalisée à l'aide des espaces et des marques de ponctuation. Par exemple :

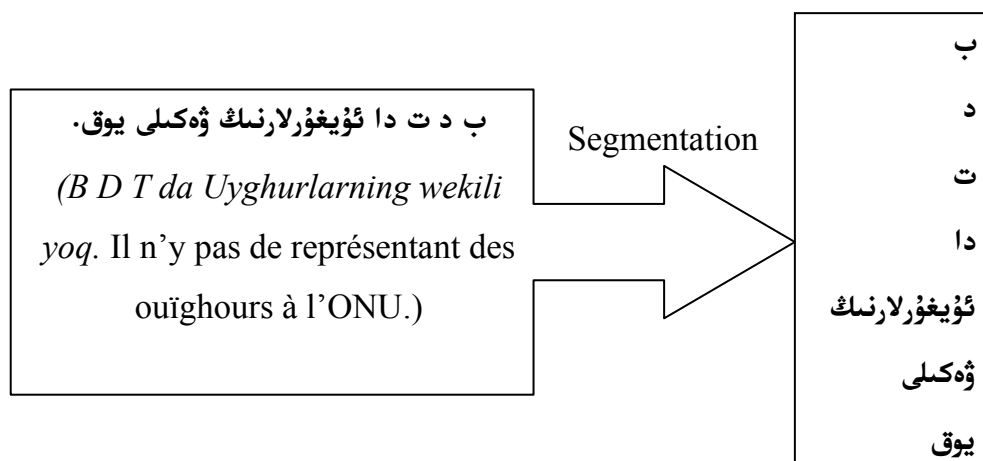


Figure 3. Un exemple de segmentation

Dans cet exemple, nous mentionnons que l'analyseur ne prend pas en compte la règle de l'abréviation ouïghoure. Autrement dit, le « *B D T* » ne devrait pas traiter comme trois mots séparés, malgré l'existence d'espaces entre les lettres. De plus, étant donné les dispositions des langues agglutinantes, l'ouïghour exige un autre niveau de segmentation plus avancé pour représenter un ensemble d'étiquettes représentant le comportement morphologique du mot. Par exemple : le mot « *Uyghurlarning* » peut être encore découper en trois unités pour montrer la racine « *Uyghur* », le nombre « *lar* » et le génitif « *- ning* ». Ainsi, le résultat d'analyse morphologique du mot « *Uyghurlarning* » est : N + PL + CAS.

Afin de résoudre ce problème de segmentation et de proposer un analyseur morphologique qui applique les règles de segmentation de l'ouïghour, nous recommandons de suivre une des deux approches suivantes :

1. L'utilisation des technologies de l'automate. Un automate est un ensemble d'états reliés par des arcs qui représentent les transitions entre états. Un automate qui encode plus d'informations et qui produit des sorties est appelé un transducteur. Ces automates peuvent être représentés par des expressions régulières et sont utilisés pour construire des règles décrivant le langage naturel depuis l'analyse morphologique des mots jusqu'à l'analyse de la phrase. Cette approche peut s'appliquer comme un supplément dans les outils existants.

2. Développement de nouvel outil. Si le résultat de la première approche n'est pas satisfaisant, on peut choisir de construire des nouveaux outils. Afin de fournir une bonne qualité de résultat, l'outil devrait être capable de faire la segmentation et d'identifier les éléments agglutinants. Nous proposons un algorithme en tenant compte des règles de segmentation et des formes fléchies. Cet algorithme a été implémenté dans un premier temps en utilisant des scripts Perl.

```

Pour tout les caractères dans la phrase
Découper les composant de la phrase à l'aide d'espace, stocker dans un tableau

Pour tout les éléments dans le tableau
  S'il y a encore des éléments à traiter alors
    Si l'élément courant est une partie d'abréviation alors
      Trouver la fin d'abréviation et constituer le mot entier ; fin
    Autrement si l'élément courant fait une partie des mots doublé alors
      Trouver la deuxième partie et constituer le mot entier ; fin
    Autrement si l'élément courant fait une partie des mots composés alors
      Trouver la deuxième partie et constituer le mot entier ; fin
    Autrement fin
  Tant Que (l'élément courant n'est pas la racine ou décomposable) faire
    décomposer les éléments agglutinants (suffixes syntactiques et dérivatifs)
  Fin de tant que
Autrement fin
Fin

```

Figure 4. Algorithme de segmentation de phrase

Après ce traitement, nous obtenons séparément toutes les formes des mots (voir figure 6), éléments agglutinants et des marques de ponctuations. Afin d'afficher le résultat d'analyse morphologique et d'associer les formes surfaces aux formes lexicales nous proposons l'algorithme ci-dessous :

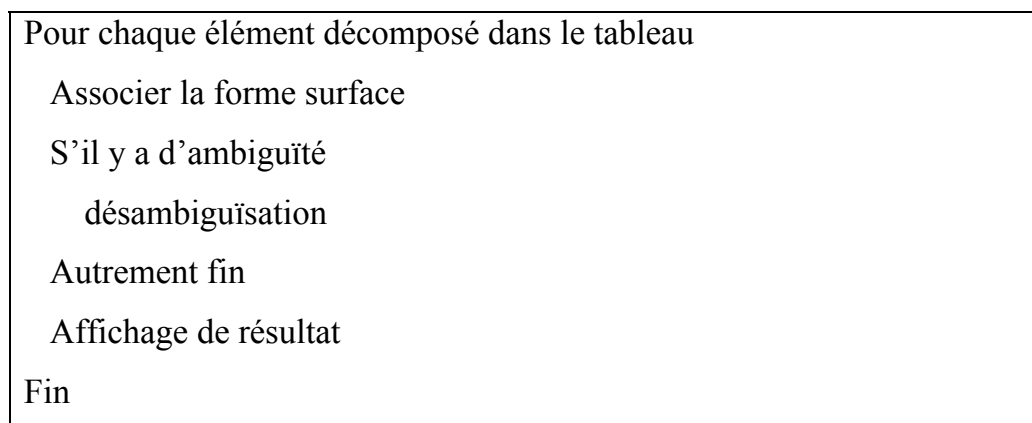


Figure 5. Algorithme d'association des formes surfaces aux formes lexicales

A la fin de ces deux processus, le système doit être capable de donner le résultat d'analyse morphologique. Par exemple :

<i>B D T</i>	N
<i>da</i>	CAS
<i>Uyghurlarning</i>	N + PL + CAS
<i>wekili</i>	N + 3POS
<i>Yoq</i>	ADJ

Figure 6. Analyse morphologique d'une phrase

Il est probable qu'un mot peut avoir des sens différents selon le contexte de son utilisation, mais un seul sens est valide lorsque le mot est dans son contexte. Ce problème exige certainement du traitement de la désambiguïsation. Dans ce cas, nous pouvons utiliser une des deux méthodes : désambiguïsation sémantique ou désambiguïsation statistique. Nous traitons seulement l'ambiguïté concernant l'extraction d'entités nommées dans le paragraphe suivant.

7.5 Extraction des entités nommées

L'extraction des entités nommées est un processus de marquage et normalisation des noms propres (personnes, locations, amorces et organisations) et des expressions structurés (valeur pécuniaire, événements, dates et temps) à partir d'un document textuel. Kemal Oflazer [Oflazer, 2001] présente une approche basée sur « *n-gram language models in*

hidden Markov models » et propose 4 modèles¹⁹⁹ (lexicale, contextuel, morphologique, insigne nominatif²⁰⁰) pour l'extraction des entités nommées turques. Dans cette étude, nous utilisons la même approche pour l'ouïghour en submergeant les deux derniers modèles.

7.5.1 Un modèle pour traiter l'ouïghour

Dans le chapitre 3 nous avons présenté la constitution des noms, noms personnels, l'ordre des suffixes de nombre, de possession et des cas. Maintenant, nous les appliquons pour capturer des entités nommées dans le graphe²⁰¹ suivant :

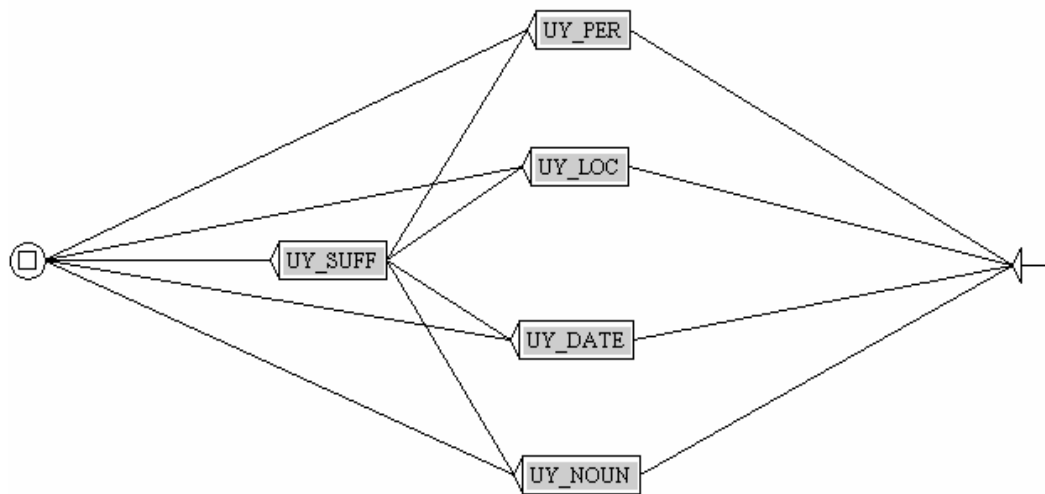


Figure 7. Un automate d'extraction des entités nommées ouïghoures

Nous voyons cinq boîtes dans cet automate. Chacune des boîtes représente un graphe indépendant qui permet d'appeler dynamiquement les graphes lexicaux pour récupérer des entités nommées. UY-PER, UY_LOC, UY_DATE et UY_NOUN correspondent respectivement aux racines lexicales de noms personnels, locations, dates et noms. UY_SUFF correspond aux suffixes (nous savons bien qu'en ouïghour les suffixes peuvent s'associer avec n'importe quelle type d'élément lexique comme : nom, verbe, adjectif). La séquence UY_XXX → UY_SUFF correspond donc aux entités nommées suffixées. Par exemple : *Warislar* (Waris etc.), *Fransiyelikler* (les français), *1-dékabirdin* (du 1 décembre), *güllüklerdin* (des jardins).

¹⁹⁹ voir Kemal Oflazer & MM Tür, 2001, « A statistical information extraction system for Turkish », page 17.

²⁰⁰ *Name Tag Model*.

²⁰¹ Les boîtes grises renvoient aux sous graphes dont elles portent les noms.

Le graphe suivant montre une partie des éléments qui se trouvent dans la boîte UY_SUFF. Il permet de récupérer une séquence valide pour l'extraction des noms (dans ce graphe, pour le possessif nous avons appliqué P1 du Tableau 13, et nous n'avons pas appliqué tous les suffixes). Si la boîte UY_NOUN envoie un nom « *Fransiye* », dans ce cas le grave ci-dessous est capable de trouver: *Fransiye* (France), *Fransiyege* (à la France), *Fransiyede* (en France), *Fransiyelik* (français), *Fransiyelikler* (français/es), *Fransiyeliklarning* (des français/es), *Fransiyeliklarningki* (des français/es, REL), *Fransiyeliklarningkige* (aux français/es), *Fransiyem* (ma France), *Fransiyemge* (à ma France), *Fransiyelikim* (mon/ma français/es), *Fransiyeliklirim* (mes français/es), *Fransiyelikimge* (à mon/ma français/es), *Fransiyeliklirimge* (à mes français/es), *Fransiyelikimning* (de mon/ma français/es), *Fransiyeliklirimning* (de mes français/es).

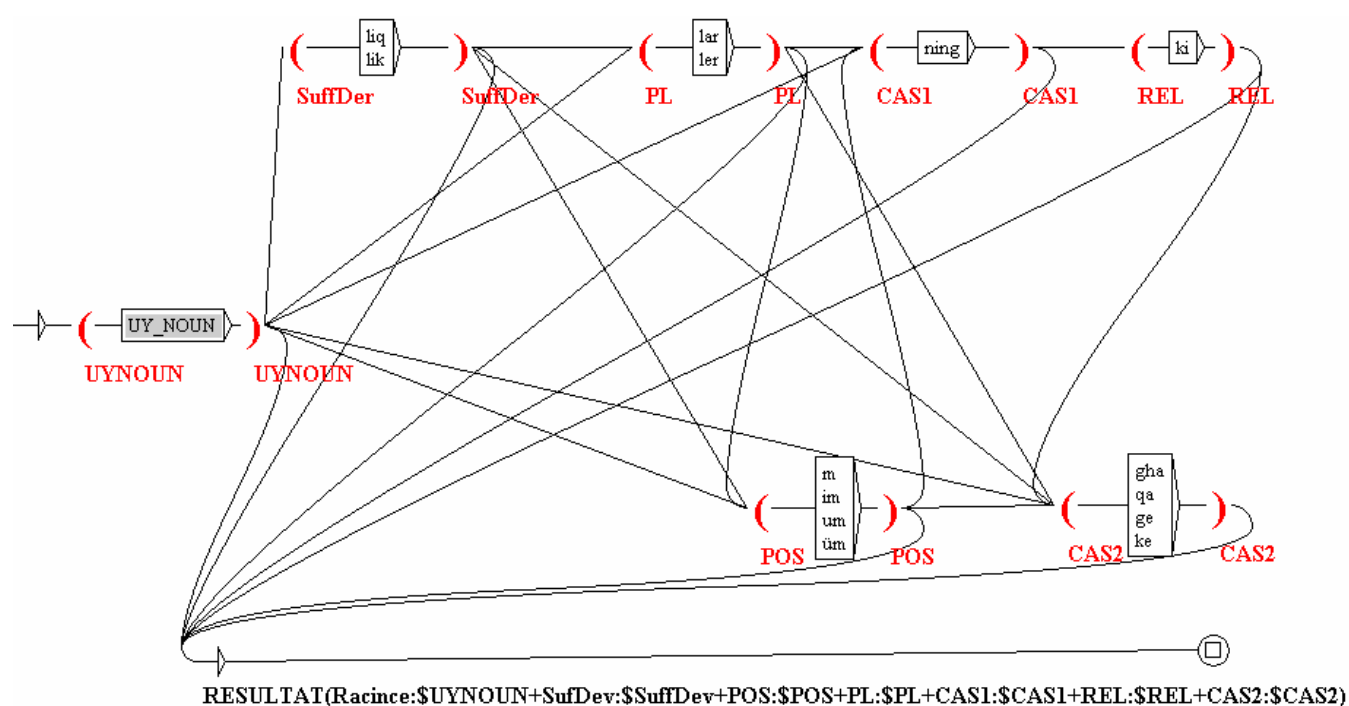


Figure 8. Un exemple d'automate pour l'extraction des noms ouïghours

Dans ce graphe, nous remarquons que la décomposition de morphèmes lexicaux en unités de base permet de leur associer une information sémantique. Après l'avoir appliqué²⁰², nous obtenons le résultat d'analyse compositionnelle en format ci-dessous :

²⁰² Nous avons appliqué en mode « Merge with input text » sur « Grammare outputs » de l'Unitex 1.0.

. . . *Fransiyeliklarningkige* RESULTAT (Racine:Fransiye +SuffDer:lik +POS:
+PL:ler +CAS1:ning +REL:ki+CAS2:ge) . . .

Il indique que les valeurs des unités sémantiques \$UYNOUN, \$SuffDer, \$POS, \$PL, \$CAS1, \$REL, \$CAS2 dans le dernier nœud du graphe sont obtenus des autres nœuds du graphe. Elles correspondent respectivement aux : racines, suffixes dérivatifs, possessifs, pluriels (le nombre), cas et relatifs. S'il n'y a rien entre les signes « : » et « + », il indique qu'il n'y a pas d'unités sémantique.

Nous mentionnons aussi que l'ordre des suffixes de nombre, de possession et des cas respecte l'ordre que nous avons présenté dans le paragraphe 3.6.2.4 . A savoir :

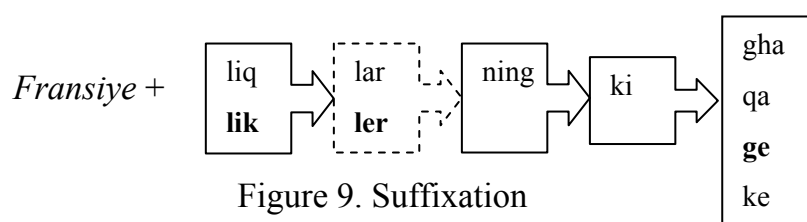
Fransiyeliklarningkige = *Fransiye* + *lik* + *ler* + *ning* + *ki* + *ge*

Racine + SuffDer + PL + CAS + REL + CAS.

7.5.2 Problèmes d'harmonie vocalique et de l'affaiblissement des voyelles

Dans le chapitre 6 nous avons présenté deux phénomènes particuliers de la langue ouïghour – harmonie vocalique et affaiblissement des voyelles. Ceux-ci existent également dans les autres langues agglutinantes comme le turc et le finnois. Marie CALBERG [CALBERG, 2003] présente un modèle pour traiter la morphologie du finnois, et elle arrive à contrôler les phénomènes d'assimilation de voyelles²⁰³. Nous appliquons un des ses graphes pour résoudre partiellement le problème de filtrage des suffixes casés par le phénomène d'harmonie vocalique.

Dans le paragraphe précédent, nous avons vu que le mot *Fransiye* doit choisir ses propres suffixes pour construire *Fransiye+lik+ler+ning+ki+ge*. C'est-à-dire, en respectant les règles d'harmonie vocalique de consonnes et de voyelles (voir paragraphe 3.2.3 et 3.2.4), l'analyseur doit être capable de choisir les suffixes composés des voyelles antérieures du tableau suivant :



²⁰³ Voir « Traitement de la morphologie du finnois par transducteur à nombre fini d'état », Marie Calberg 2003.

Marie CALBERG (2003) a proposé, selon l'analyse de Clémenceau (1996), ce modèle permet de modifier dynamiquement les cas de mise en relation de racines avec des suffixes. Nous le présentons avec un peu de modification pour adapter à la morphologie de l'ouïghour.

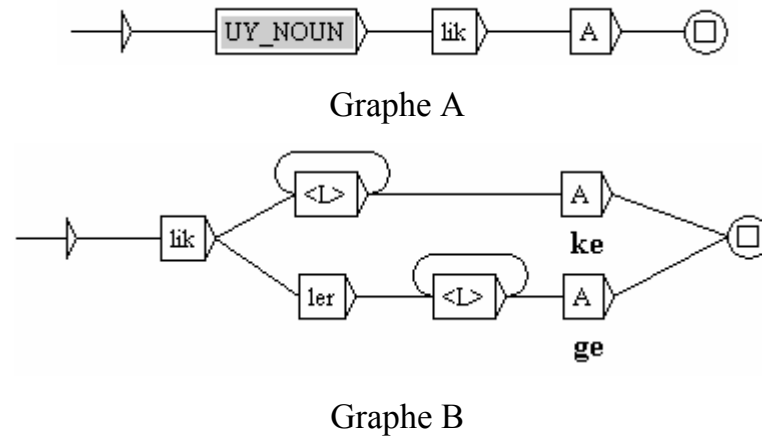


Figure 10. Harmonie Vocalique

Le graphe B doit être appliqué après le graphe A afin de sélectionner un suffixe adéquat. Les deux graphes indiquent le fait que la variable A doit être réalisée par *ke* si et seulement si elle est précédée par un suffixe de voyelle médiane *lik* ; par *ge* si et seulement si elle est précédée par un suffixe de voyelle médiane *lik* suivi d'un suffixe de voyelle antérieure. Le <L> correspond à n'importe quelle lettre de l'alphabet. Nous pouvons générer facilement les graphes correspondants aux suffixes des voyelles médianes postérieures selon cette base, nous ne les présenterons pas dans cette étude.

7.5.3 Algorithme

Nous avons dit que le modèle ci-dessus peut résoudre partiellement le problème de filtrage des suffixes puisqu'il n'est pas applicable pour filtrer tous les suffixes. Ainsi, il n'est pas possible de trouver toutes les formes lexicales valides en utilisant les règles d'harmonie vocalique. Afin d'avoir un résultat satisfaisant et de résoudre définitivement les problèmes d'harmonie vocalique, d'affaiblissement des voyelles et de filtrage des suffixes, nous proposons l'algorithme ci-dessous. Il sert à la désambiguïsation et l'identification des formes lexicales pendant le processus de l'extraction de l'information.

pour chacun des mots dans le « sac de noms »
segmenter en syllabe
récupérer la voyelle et la dernière consonne de la dernière syllabe
si la voyelle est une voyelle affaiblie alors
trouver la voyelle originale
autrement si la dernière voyelle est une voyelle médiane alors
récupérer la voyelle précédente; fin
tant que (il y a des suffixes à associer)
filtrer les suffixes en utilisant l'arbre Janbaz
choisir les suffixes adéquats
associer des suffixes au mot et créer des nouveaux mots
modifier la forme surface
fin de tant que
Fin

Figure 11. Algorithme de génération des entités nommées et de désambiguïsation

Dans cet algorithme, le « sac de noms » indique un ensemble d'entités nommées non suffixées. Cet algorithme permet de produire des nouveaux mots valides, en choisissant le format adéquat parmi 247 suffixes dérivatifs et plus de 320 suffixes syntactiques, à partir du « sac de noms », il peut générer la forme surface et la stocker dans un fichier afin de visualiser le résultat du traitement morphologique.

7.6 Evaluation

Cette évaluation a été effectuée sur un document de 5330 mots que nous avons élaboré pour tester les automates. Nous avons manuellement fait la segmentation et nous avons appliqué la méthode proposée dans ce chapitre pour souligner des entités nommées — les dates, les noms personnels et les noms des pays. Nous avons réussi à récupérer les formes surfaces (racine, suffixe de dérivation, possessif, cas etc.) des mots existant dans le corpus.

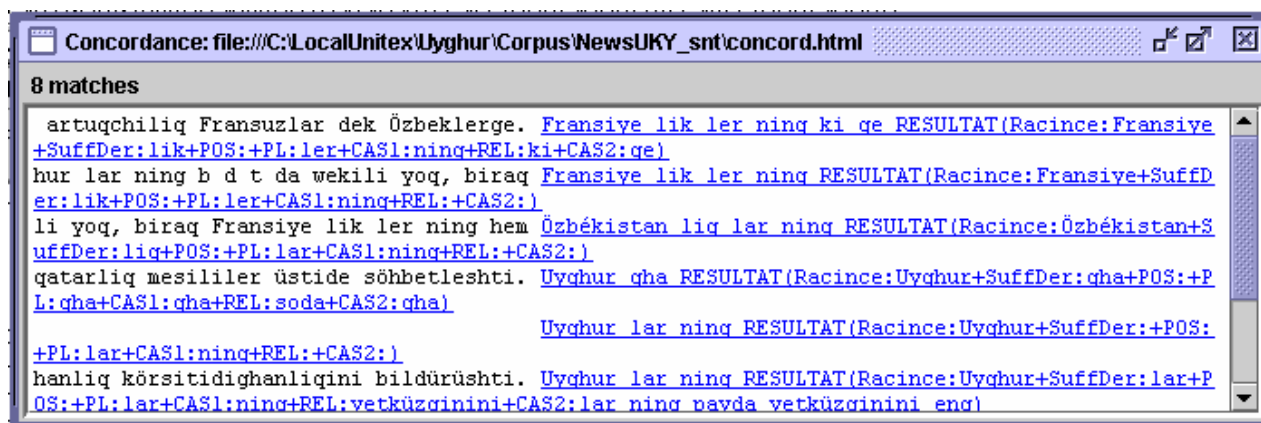


Figure 12. Résultat d'analyse

L'analyse statistique pour tester la précision et le rappel n'a pas été faite en raison de l'absence d'un large corpus d'évaluation. L'Unitex ne nous permet pas de décomposer dynamiquement des entrées pour produire une analyse morphologique. Il ne peut pas non plus résoudre les problèmes d'affaiblissement des voyelles que nous avons présenté dans le paragraphe 3.2.2. car le phénomène s'effectue dans la racine et aussi dans le suffixe. Nous envisageons aussi d'appliquer cette méthode à d'autres domaines et de réfléchir comment l'information linguistique nécessaire pour mener à bien cette tâche peut être acquise de manière partiellement automatique afin de minimiser le temps de développement d'un nouvel outil pour le traitement de l'ouïghour.

Chapitre 8. Parseur et correcteur d'orthographe morphologique

8.1 Bref état des recherches

De nombreux projets de recherches²⁰⁴ ont été menés, à l'Université du Xinjiang, dans le cadre du correcteur d'orthographe ouïghour, notamment par M. Ablimit, K. Ubul, G. Adongbieke. Plusieurs logiciels de correcteur d'orthographe ont été développés par des entreprises privés et des chercheurs indépendants, à savoir: *Kenji Correcteur*²⁰⁵, *Yulghun Editor*²⁰⁶, *Udmish Correcteur*²⁰⁷, *Uighursoft Correcteur*²⁰⁸, *Irpan Correcteur*²⁰⁹. Les deux premiers sont les logiciels libres les plus utilisés parmi les internautes ouïghours, *UyghurEdit* (G. T. Kenji, 2008) et *Yulghun Editor* (M. Abdullah, 2006) recensent 632 000 et 247 847 entrées respectivement. Tous les logiciels de correcteur d'orthographe mentionnés ci-dessus utilisent une approche basée sur la notion de «corpusage»²¹⁰. Ils considèrent tous les mots qui ne sont pas dans leurs listes comme des fautes d'orthographe. Voici un extrait de la liste des mots dérivés de la forme affaiblie (كېل) du verbe primitif «كەل» (venir) proposée dans *Yulghun Editor*:

كېلىۋاتسام	كېلىۋاتسىمۇ	كېلىۋاتقان	كېلىۋاتقاندا	كېلىۋاتقاندەك	...
كېلىۋاتقانلاردىن	كېلىۋاتقانلارغا	كېلىۋاتقانلارنىڭ	كېلىۋاتقانلىرى	كېلىۋاتقانلىرىدىن	كېلىۋاتقانلىقى
كېلىۋاتقانلىقىدىن	كېلىۋاتقانلىقىغا	كېلىۋاتقانلىقىم	كېلىۋاتقانلىقىمنى	كېلىۋاتقانلىقىمىز	كېلىۋاتقانلىقىنى
كېلىۋاتقانلىقىنىڭ	كېلىۋاتقانلىقىدى	كېلىۋاتقۇدەك	كېلىۋاتقۇلى	كېلىۋاتقۇلىمۇ	كېلىۋاتقۇلىنى
كېلىۋاتقۇلىنىڭ	كېلىۋاتقۇلىدى	كېلىۋاتقۇلىدەك	كېلىۋاتقۇلىلىرى	كېلىۋاتقۇلىلىرىدىن	كېلىۋاتقۇلىلىقى

²⁰⁴ Voir <http://www.ilib.cn/A-xtgcllysj200305021.html>, «Research on Uighur Corrector System in Multilingual Environment», Mijit Ablimit & Kurban Ubul, 2003; <http://www.ilib.cn/A-zwxxxb200406009.html>, «Research on Uighur Word Segmentation», Gulila Adongbieke, Mijit Ablimit, 2004; projet financé par «National Natural Science Foundation of China», no. 69963002. Nous n'avons pas pu tester l'outil proposé en raison de sa non disponibilité sur le marché.

²⁰⁵ Voir <http://kenjisoft.homelinux.com/uyghuredit> pour la dernière actualité sur la mise à jour de la liste des mots.

²⁰⁶ Voir <http://www.yulghun.com/news/editor16.html>

²⁰⁷ Voir <http://udmishkorrektor.googlepages.com/index.html>

²⁰⁸ Voir <http://www.uighursoft.com/uighur/ProductShow.asp?id=8>

²⁰⁹ Voir <http://www.irpan.com/Web/Forum/FTopicView.aspx?TID=58>

²¹⁰ Le fait d'ajouter toutes les formes des mots dans un corpus, appelé dictionnaire de formes.

كېلىۋاتقىنىمىزنى	كېلىۋاتقىنىمىزغا	كېلىۋاتقىنىمىزدا	كېلىۋاتقىنىمىز	كېلىۋاتقىنىمىنى	كېلىۋاتقىنىمىغا
كېلىۋاتىدىكەنە	كېلىۋاتىدۇ	كېلىۋاتمايدۇ	كېلىۋاتمايتتى	كېلىۋاتماقتا	كېلىۋاتقىنىنى
كېلىۋاتىمىز	كېلىۋاتىمەنغۇ	كېلىۋاتىمەن	كېلىۋاتىسىز	كېلىۋاتىسەن	كېلىۋاتىدىلا
كېلىۋالدىم	كېلىۋالدى	كېلىۋالدىق	كېلىۋالايلى	كېلىۋالائىتى	كېلىۋال
كېلىۋالغانلىقىنى	كېلىۋالغاندىن	كېلىۋالغاندى	كېلىۋالغاندەك	كېلىۋالغان	كېلىۋالغۇن
كېلىۋالدىغان	كېلىۋالدىۇ	كېلىۋالماقچى	كېلىۋالماق	كېلىۋالغىلى	كېلىۋالغانىدى
كېلىۋەرمەك	كېلىۋەرگەن	كېلىۋەرسە	كېلىۋەردىم	كېلىۋەردى	كېلىۋەر
كېلىۋېرىدۇق	كېلىۋېتىپلا	كېلىۋېتىپتۇ	كېلىۋېتىپ	كېلىۋەرمەيدۇ	كېلىۋەرمەي
كېلىۋالايلى	كېلىۋالائىتى	كېلىۋېدىم	كېلىۋېدى	كېلىۋاتقانلار	...

Table 8.1 Extrait d'entrée du correcteur de *Yulghun Editor*

Dans la table précédente on constate que si «كېلىۋاتقانلاردىن» existe, la forme suivie de l'interrogatif «مۇ» est absente. De ce fait, les deux logiciels considèrent «كېلىۋاتقانلاردىنمۇ» comme une faute d'orthographe, bien que ce ne soit pas le cas. L'introduction d'un nouveau mot nécessiterait d'ajout d'une entrée dans le corpus. Dans le 5. 9.1 nous avons montré la possibilité de créer 24 294 nouveaux mots à partir de la racine d'un seul verbe. Si on multiplie ce chiffre par le nombre des verbes primitifs²¹¹, on obtient le nombre de 15 135 162, qui représente le total théorique des formes verbales. Comme ce chiffre ne couvre que les verbes, on peut aisément imaginer la taille d'un corpus regroupant toutes les dérivations possibles de toutes les catégories lexicales, alors qu'un dictionnaire de français courant, par comparaison, contient autour de 50 000 entrées ou lemmes. Pour une langue agglutinante comme l'ouïghour, établir une liste complète des mots et de toutes leurs variantes possibles apparaît donc comme une tâche pratiquement infaisable. Dans ce chapitre, nous étudions une méthode qui peut réduire la masse de données dans le corpus tout en couvrant un maximum de possibilités lexicales.

²¹¹ Nous avons recensé 623 verbes primitifs (ou racines verbales non dérivées) en 2007. Ce chiffre initial pourrait augmenter.

8.2 Qu'est ce qu'un parseur ?

La définition de parseur que nous utilisons est celle de Dick Grune et Ceriel J.H. Jacobs²¹², qui ont fait un résumé global qui couvre un large assortiment de textes.

“Parsing is the process of structuring a linear representation in accordance with a given grammar. This definition has been kept abstract on purpose, to allow as wide an interpretation as possible. The “linear representation” may be a sentence, a computer program, a knitting pattern, a sequence of geological strata, a piece of music, actions in ritual behaviour, in short any linear sequence in which the preceding elements in some way restrict the next element.” [Grune et Jacobs, 1990, p.13]

Au sens linguistique, un parseur est donc un outil informatique qui permet de faire une analyse syntaxique d'un texte, d'une entité (groupe de mots) d'un mot ou d'une chaîne de caractères pour obtenir une représentation linéaire de l'objet en question. Il divise l'objet d'analyse en petite unité fonctionnelle et il peut expliquer les relations morpho-syntaxiques entre les unités individuelles. Quand l'objet d'analyse est un mot dans langue naturelle, un parseur doit pouvoir trouver la racine et les suffixes qui le suivent et il doit pouvoir analyser les ambiguïtés structurelles potentielles, tout en respectant les règles de la grammaire de la langue dans laquelle le mot est écrit. Prenons deux exemples :

1) 0033148064684 (un numéro de téléphone)

2) *Küchlendürelmeywatqanliringlarningkidinmu?*

(« Sont-ils parmi ceux que vous êtes incapable de renforcer ? »)

Pour le premier, un parseur doit être capable de dire : a) « 0033 » est un code identifiant de pays – la France - sachant qu'il est aussi possible d'utiliser +33 ; b) le cinquième chiffre « 1 » est équivalent du 01 utilisé lorsque le numéro est composé en France, c'est-à-dire sans le préfixe international et constitue le code de la région Ile-de-France ; c) le « 48 » est dédié au quartier ou à l'arrondissement X . Le parseur doit pouvoir

²¹² Parsing Techniques - A Practical Guide, England, 1990; ISBN 0 13 651431 6

également traiter, en tant que cas spéciaux, les chiffres séparés par un ou plusieurs traits d'union «-» ou regroupés par des parenthèses « () », comme dans (613) 555-1234.

Le deuxième correspond à une phrase exprimée en ouïghour par une seule racine lexicale suivie d'une série de suffixes. Un parseur ouïghour doit pouvoir identifier la racine – « *küch* » et chacun des suffixes qui la suivent avec leur représentations morpho-syntactiques. Les suffixes devront être découpés en le+n+dür+el+mey+wat+qan+lir+ing+lar+ning +ki+din+mu (le signe + indique la frontière entre deux suffixes. Voir le chapitre 6 pour plus de détails sur les informations sémantiques et les règlements linguistiques.) Dans les paragraphes suivants, nous expliquons les spécificités d'un parseur ouïghour.

8.3 Prétraitement

La première opération qu'un parseur effectue sur l'objet de traitement est sa fragmentation en entités distinctes. Un objet de traitement est une chaîne de caractères initiaux qui peut être, par ordre de volumétrie décroissante, un texte, un paragraphe, une phrase ou une entité. Une entité est une unité de taille supérieure à celle d'un mot simple. Elle peut être un mot composé ou une expression composée de plusieurs mots. Nous considérons comme mots les entités qu'un programme est capable d'identifier. Avant d'entrer dans la procédure d'identification des entités distinctes, il est important de faire un « nettoyage » afin de diminuer la charge de traitement. Nous étudions ce prétraitement en trois étapes.

8.3.1 La segmentation en mots

Dans une phrase, on trouve des mots, des espaces, des chiffres et des signes de ponctuation. Au sens strict, on peut dire que la segmentation est une procédure de découpage en mots d'une phrase. La segmentation peut être réalisée à l'aide de séparateurs :

ponctuation : espace, tabulation, retour chariot, point, virgule, deux-points, point d'interrogation, point d'exclamation, parenthèses, chiffres, trait d'union, tiret cadratin et des guillemets.

caractères spéciaux : symboles et lettres qui n'existent pas en ULY, UEY et USY.

suffixes précédés d'un mot emprunté écrit en langue d'origine. Ex. :

1) مېكروسوفت ۋوردتىكى ئالاھىدە ماھارەتلەرنى ئىنتېرنېتتىن تاپقىلى بولىدۇ

2) Microsoft Word تىكى ئالاھىدە ماھارەتلەرنى Internet بولىدۇ

Les deux phrases ci-dessus sont considérées comme correctes, malgré le fait que, dans la première, les termes «Microsoft Word» et «Internet» sont transcrits lettre-par-lettre en UEY et dans la deuxième ils sont écrits en leur langue d'origine. En ouïghour, comme il n'y a pas de stricte obligation de transcrire les noms propres empruntés à des langues étrangères, ce genre d'utilisation est toléré.

ئۇيغۇر ئېلىپبەسىدىكى ھەرپلەر ش ئۇ ئا ر مىللەتلەر تىل-يېزىق خىزمىتى كومىتېتى ۋە شىنجاڭ ئۇنىۋېرسىتېتىدىكى تەتقىقاتچىلارنىڭ تىرىشچانلىقى بىلەن 1998-يىلى خەلقئارالىق ISO 10646 ئۇچۇر بىر تەرەپ قىلىش ئۆلچىمىگە كىرگەن ھەمدە خەلقئارا بىردىنبىر كود جەدۋىلى Unicode قا كىرگۈزۈلگەن. بىراق، تا ھازىرغىچە ئۇيغۇرچە يۇمشاق دېتاللار تېخىچە ئاپتونوم رايونىمىزدىكى ھەر دەرىجىلىك (نازارەت، ۋىلايەت، شەھەر، ناھىيە، يېزا) ھۆكۈمەت ئورگانلىرى، دۆلەت كادىرلىرىنىڭ ئىشخانىلىرى، ئالىي ۋە ئادەتتىكى مائارىپتىكى ئوقۇتقۇچىلار ئىشخانىلىرى، كەسپىي ۋە مۇلازىمەت ئورۇنلىرىنىڭ كومپيۇتېرلاشقان ئىشخانىلىرىدا تولۇق ئومۇملاشقنى يوق.

(Extrait d'un texte ouïghour avant la segmentation)

ئۇ	ش	ھەرپلەر	ئېلىپبەسىدىكى	ئۇيغۇر
يېزىق	تىل	مىللەتلەر	ر	ئا
ئۇنىۋېرسىتېتىدىكى	شىنجاڭ	ۋە	كومىتېتى	خىزمىتى
يىلى	1998	بىلەن	تىرىشچانلىقى	تەتقىقاتچىلارنىڭ
بىر	ئۇچۇر	ISO	10646	خەلقئارالىق
ھەمدە	كىرگەن	ئۆلچىمىگە	قىلىش	تەرەپ
Unicode	جەدۋىلى	كود	بىردىنبىر	خەلقئارا
ھازىرغىچە	تا	بىراق	كىرگۈزۈلگەن	قا
ئاپتونوم	تېخىچە	دېتاللار	يۇمشاق	ئۇيغۇرچە
ۋىلايەت	نازارەت	دەرىجىلىك	ھەر	رايونىمىزدىكى
ئورگانلىرى	ھۆكۈمەت	يېزا	ناھىيە	شەھەر
ۋە	ئالىي	ئىشخانىلىرى	كادىرلىرىنىڭ	دۆلەت
كەسپىي	ئىشخانىلىرى	ئوقۇتقۇچىلار	مائارىپتىكى	ئادەتتىكى
ئىشخانىلىرىدا	كومپيۇتېرلاشقان	ئورۇنلىرىنىڭ	مۇلازىمەت	ۋە
		يوق	ئومۇملاشقنى	تولۇق

Table 8.2 liste des entités après segmentation

Après avoir obtenu la liste des mots segmentés, chacun de ces mots sera envoyé à la procédure qui vérifie l'orthographe (voir la figure 8.4).

8.3.2 Reconnaissance des écritures : ULY, UEY, USY

La plupart des textes sont écrits dans un seul système d'écriture. Cependant, il n'est pas rare que des mots ou des phrases écrites dans un autre système d'écriture soient insérés dans le texte. Dans l'extrait du texte ouïghour ci-dessus, on remarque que les deux termes « ISO, Unicode » apparaissent en caractères latins. Pour ce texte, un outil de correcteur d'orthographe ouïghour ne doit pas les considérer comme des mots ouïghours. Afin d'effectuer la correction orthographique d'un texte ouïghour ou des mots ouïghours dans un texte multi-écriture, l'outil doit pouvoir identifier si le mot en cours de vérification est un mot ouïghour. Étant donné que la langue ouïghoure utilise trois systèmes d'écriture (voir 3 pour plus de détails), il faut reconnaître le système d'écriture utilisé avant d'entrer dans la procédure de vérification et ne retenir pour l'étape parsing que les mots segmentés qui ne contiennent pas de caractères hétérogènes à ce système.

- UEY : l'alphabet UEY est composé de 32 lettres. Les voyelles en forme voyelle-conjointe sont précédées par une 33^{ème} lettre, «²¹³» , utilisée uniquement dans ce contexte (voir annexe 7 pour les lettres et leurs codes). La notation de toutes les voyelles et la règle d'utilisation de cette 33^{ème} lettre sont avec l'absence de certaines lettres figurant dans l'alphabet arabe classique deux caractéristiques qui permettent de distinguer les mots ouïghours des mots d'autres langues écrites dans un système graphique basé sur l'arabe. Si la valeur Unicode de chacune des lettres d'un mot se trouve parmi les valeurs des 33 lettres UEY, on considère que c'est une chaîne de lettres ouïghoures « légales » qui peut être envoyée à la procédure de vérification d'orthographe. Dans le cas contraire, le mot est traité comme une faute d'orthographe. Par

²¹³ ARABIC LETTER YEH WITH HAMZA ABOVE

exemple : ²¹⁴ يېڭى، ئىلىغىزغا، مۇبارەك، جىغرافىيا : mots dans lesquels toutes les lettres figurent dans l'alphabet UEY, sont fort susceptible d'être des mots ouïghours mais ceci ne veut pas dire à ce stade qu'ils sont des mots sans fautes d'orthographe. Par contre, خىوان، اللہ، العربیة، ئالدامچى x seront éliminés parce qu'ils contiennent des lettres qui ne figurent pas dans l'UEY²¹⁵. Dans cette situation, ils seront traités comme des fautes d'orthographe et passeront directement à la procédure de suggestion.

- ULY : il utilise 25 lettres de l'alphabet latin²¹⁶, trois lettres diacritiques (ö, ü, é) et l'apostrophe. La reconnaissance automatique des mots ouïghours écrits en ULY présente une difficulté particulière dans un texte multilingue écrit en alphabet latin. Par exemple :

*Men In'glizche «how do you do» din bashqa gep
bilmeymen*²¹⁷.

[je anglais «how do you do» à part mot ne sais pas].

L'identification de la langue et de l'écriture de chaque mot de cette phrase est une procédure lente et délicate. Comment peut-on savoir que «men, din » sont des mots ouïghours et non pas en anglais ? Ou vice-versa ? Faut-il chercher ces mots d'abord dans un dictionnaire ouïghour puis dans un autre dictionnaire anglais ? Faut-il utiliser les guillemets comme indice ? Comment faire en cas d'absence de tels marqueurs ? Est-ce que la structure syllabique

²¹⁴ Les premier et troisième mots, *yéngi* (nouveau) et *mubarek* (béné) sont ouïghours ; le second *jilingiz* (votre année) est une forme dialectale de *yilingiz* dans laquelle n'apparaît aucune lettre hétérogène à UEY ; le quatrième mot est un mot arabe qui signifie géographie mais dont l'écriture fait appel à des lettres qui sont communes à l'alphabet arabe classique et au système UEY.

²¹⁵ Le premier, le mot arabe حيوان (animal) utilise une lettre, le ح absente de UEY. Dans le second qui est la graphie arabe du nom de Dieu, *Allah*, la voyelle initiale «ا» n'est pas précédée du «ئ» demandé par les règles de l'UEY et des signes de notation vocalique et de redoublement consonantiques purement arabes sont utilisés. Le troisième, العربیة (l'arabe) utilise deux lettres arabes absentes de UEY. Quant au quatrième, il introduit le caractère latin x.

²¹⁶ La lettre « v » n'est pas utilisée et la lettre « c » n'apparaît que dans la combinaison «ch».

²¹⁷ Traduction : *je ne sais que dire «how do you do » en anglais.*

des mots ouïghours peut contribuer à l'identification ? Notre objectif ultime n'est pas de développer un outil de reconnaissance automatique des langues, mais d'étudier la possibilité de faire un correcteur d'orthographe en utilisant une approche linguistique. Nous considérons donc tous les mots qui ne contiennent que des lettres de l'alphabet ULY comme des mots ouïghours et les envoyons à la procédure de vérification d'orthographe. Les mots comme « *can, jack, virus, Bänke* » seront donc éliminés au stade de la reconnaissance des écritures mais tel ne sera pas le cas pour *how, do, you* et *do*, qui seront traités comme des fautes d'orthographe et soumis à la procédure de suggestion.

- USY : il utilise 38 lettres cyrilliques²¹⁸ et entre dans le même cas que l'ULY quand il s'agit de reconnaissance automatique des mots ouïghours dans un texte écrit en cyrillique.
- autre : les mots écrits dans une écriture autre qu'UEY, ULY et USY ne font pas l'objet de traitement supplémentaire pour la vérification d'orthographe.

On remarque que dans un texte multilingue (ou multi-écriture) en écriture latine ou cyrillique, tous les mots passent par une procédure de vérification d'écriture, quelle que soit la langue dans laquelle ils sont écrits. Comme la procédure de vérification d'orthographe²¹⁹ se fait sur la base de l'écriture UEY, une translittération vers ce système de tous les mots retenus en ULY et USY est nécessaire. La figure 1 ci-dessous présente un résumé schématique de l'ensemble de l'étape du prétraitement.

²¹⁸ Qui contient également les 6 lettres ы ё ц э ю я utilisées uniquement dans les mots d'emprunt récent au russe. Voir Annexe 10.

²¹⁹ Ce module sera présenté en détail dans les sections suivantes.

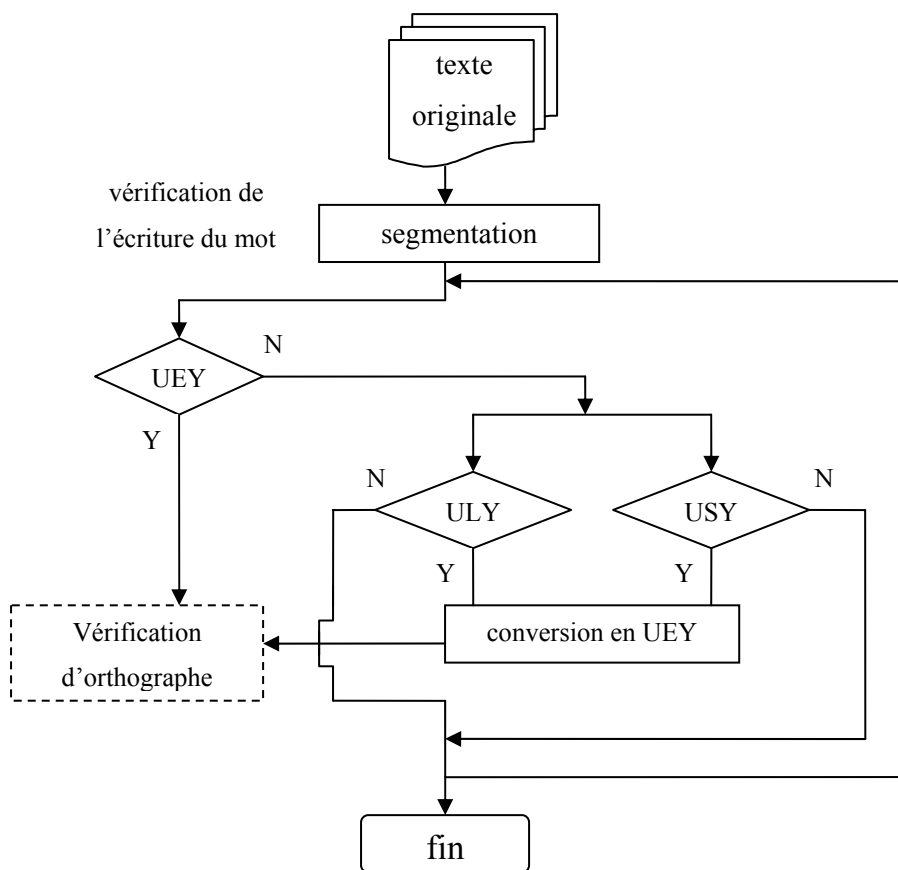


Figure 8.1 Procédure de prétraitement

8.4 Identification de la racine

Un parseur doit pouvoir déterminer si le mot en cours d'analyse est dérivé d'un autre mot, et si c'est le cas, comment ce mot est dérivé. Afin répondre à la première question, il faut trouver la racine. Pour ce faire, le parseur doit utiliser une liste des mots (le « dictionnaire ») qui ne contient que des mots correctement orthographiés. Cette liste ne contient que des racines primitives et les formes particulières des mots irréguliers²²⁰. Nous avons recensé²²¹ 617 verbes dans notre liste préliminaire et ils sont stockés par ordre alphabétique afin d'accélérer la recherche. La racine du mot est recherchée dans le dictionnaire en utilisant un algorithme d'appariement maximal. Dans cet algorithme, on

²²⁰ La langue ouïghoure ne comprend que très peu d'exceptions. Dont la plus notable est le mot *su* (eau), qui devient *süyü* à la forme définie, avec changement de la voyelle arrière (*u*) du radical en une voyelle d'avant (*ü*). Les autres exceptions sont toutes des racines verbales primitives affectées par le phénomène de réduction vocalique.

²²¹ Voir chapitre 5 pour savoir comment ils ont été sélectionnés.

recherche d'abord le mot entier dans le dictionnaire ; si le mot s'y trouve, il est considéré comme un mot sans suffixe donc qu'il n'est pas nécessaire de parser davantage (ex : كحل, viens). Sinon, l'une des deux méthodes de recherche suivantes sera appliquée :

8.4.1 Recherche progressive

La recherche progressive consiste à trouver la racine de l'entrée à partir d'une partie des caractères les plus à gauche, en appliquant des règles de grammaire. Les verbes primitifs les plus courts se composent généralement de trois caractères dans l'écriture UEY²²². Le point de départ est donc les trois caractères initiaux, c'est-à-dire situés le plus à gauche (marqués A sur la figure ci-dessous). Si A est retrouvé dans le dictionnaire, le reste des caractères est considéré comme un suffixe ou une suite de suffixes ; sinon, A est analysé pour déterminer s'il s'agit d'une racine modifiée par affaiblissement vocalique ou élision de consonne. Si la racine n'est pas retrouvée, A est remplacé par une suite de caractères de longueur A+1 (soit A=A+B/C/D/E...Z), avec répétition de cette procédure jusqu'à l'identification de la racine. Si la racine n'est toujours pas identifiée lorsque A+B/...Z correspond à la longueur de l'entrée, la structure de l'entrée (mot en traitement) est considérée comme incorrecte.

²²² Certains verbes primitifs avec voyelle initiale (*al, at, ur, uch, etc.*) ne présentent que deux caractères en ULY et en USY mais l'ajout obligatoire de la 33^{ème} lettre devant toute voyelle initiale en UEY les rend trilitères dans le dictionnaire de ce système basé sur le système UEY. Les trois seules exceptions à la règle sont les 3 verbes en deux lettres *de* (dire), *ye* (manger) et *yu* (laver) qui, à l'exception du dernier, ne nécessitent de traitement particulier au niveau du dictionnaire.

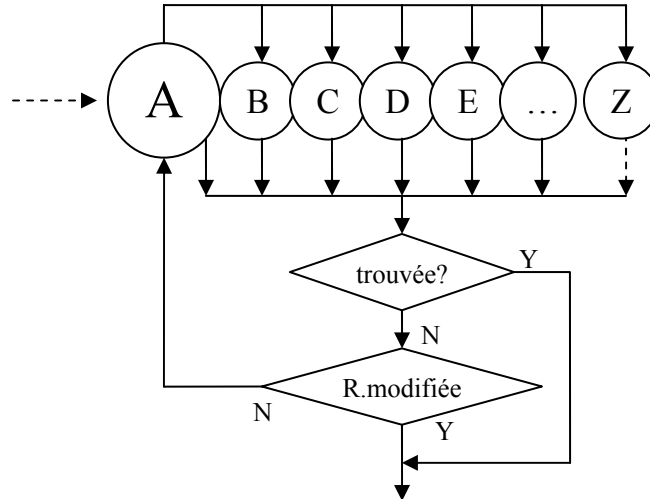


Figure 8.2 Recherche progressive

Par exemple, dans le mot “کبلی و اتقانلار”(ceux qui sont en train de venir), les trois premiers caractères A = “کب” (forme affaiblie du verbe primitif “کحل”). L’algorithme peut considérer A comme la racine puisque quand il arrive à la deuxième condition (est-ce une racine modifiée ?) la réponse est oui. Nous verrons la procédure de vérification de racine modifiée dans la section suivante.

8.4.2 Recherche dégressive

C’est l’inverse de la recherche progressive. Un premier caractère est enlevé à la droite de la chaîne de caractères de l’entrée avant le début d’une nouvelle recherche dans le dictionnaire. Si cette recherche n’est pas fructueuse, l’analyse se poursuit pour déterminer s’il s’agit d’une racine modifiée. Si ces deux recherches ne permettent toujours pas de retrouver la racine, un deuxième caractère est enlevé à droite et la même procédure se répète jusqu’à l’identification de la racine. Une fois la racine identifiée, les caractères enlevés à sa droite sont traités comme un suffixe ou une suite de suffixes. Si la racine n’est toujours pas identifiée lorsque le nombre des caractères de la séquence envoyée à la recherche atteint trois (soit A), la structure de l’entrée est considérée comme incorrecte.

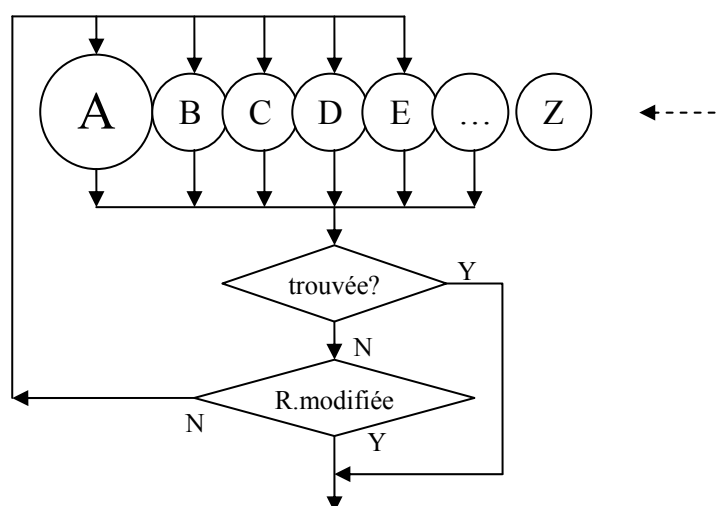


Figure 8.3 Recherche dégressive

La recherche dégressive n'est pas la procédure optimale pour trouver la racine des mots hautement suffixés parce que l'algorithme de recherche utilise beaucoup plus de temps que dans la recherche progressive pour arriver jusqu'à la racine.

8.4.3 Analyse de séquences modifiées

Deux phénomènes de transformation doivent être pris en compte ²²³ : 1) affaiblissement vocalique ; 2) élision de consonne. Le premier, qui se manifeste par la réduction des voyelles *a* et *e* en *i* ou en *e* dans certains contextes peut affecter la racine aussi bien que les suffixes. Le second n'affecte que les verbes primitifs se terminant par la consonne *l* lorsqu'ils sont suivis du suffixe *-ip/up*, lui-même réduit à un simple *-p*. Lorsque la consonne élidée *l* est rétablie à la fin du verbe primitif, la règle d'affaiblissement vocalique s'applique de nouveau et, conformément à celle-ci, le *a* ou le *e* qui la précède est affaibli en *é*, ce qui donne, avec le suffixe *-ip* la même forme *-élip* dans les deux cas. Ils peuvent être détectés grâce aux algorithmes suivants :

²²³ Voir 5.3 Voyelles et 5.8 Règles de suffixation et cas spéciaux pour les détails.

Début

extraire les voyelles de l'entrée

initialisation à vide des deux formes affaiblies possibles

SI le nombre des voyelles n'est pas zéro ALORS

 récupérer la dernière voyelle

 SI la dernière voyelle est «i» ALORS

 récupérer l'index de la dernière voyelle dans l'entrée

 SI le nombre des voyelles est supérieur à un ALORS

 remplacement de la dernière voyelle par «a» et «e», et

 alimentation des deux formes originales (non affaiblies)

possibles

Algorithme 1. Affaiblissement des voyelles ouïghoures

Début

extraire les voyelles de l'entrée

initialisation à vide des deux formes modifiées possibles

initialisation des suffixes qui résultent de l'élision (SE) → SE // -ap, -ep, -ip, -op

SI le nombre des voyelles est égal à un et SE existe dans l'entrée ALORS

 remplacement de -p par -l et alimentation de la première forme

 restitution des formes originales en remplaçant -ap/-ep/-ip/-op par -élip/-élip/-
 ilip/

 -olup respectivement et alimentation de la deuxième forme

 retourner deux formes modifiées possibles suite à l'élision

SINON ne rien faire et de retourner deux formes vide

Fin

Algorithme 2. Élision de consonne ouïghoure

Lors de l'identification de la racine, la recherche est faite soit a) en concaténant séquentiellement la racine trilitère tirée de la liste avec le premier caractère situé à sa droite (approche progressive) ; soit b) en abandonnant chaque fois le dernier caractère de la racine hypothétique (approche dégressive). Dans le deuxième cas, la recherche de racine est plus lourde pour des mots composés de plusieurs suffixes. Par exemple : «كېلىۋاتقانلارنىڭمۇ»

(*kéliwatqanlarningmu*, est-ce à ceux qui sont en train de venir ?) est composé de 18 caractères (dont la racine modifiée est *kél* $\leftarrow$ *kel*), suivie de 15 caractères : *i-w-a-t-q-a-n-l-a-r-n-i-n-g-m-u*. Avec la méthode b), l'algorithme doit exécuter la boucle (voir figure 9.2) 15 fois pour trouver la racine.

L'approche syllabique propose une solution plus rapide que l'approche décrite ci-dessus : l'entrée est maintenant composée des 7 syllabes *ké-li-wat-qan-lar-ning-mu*. Si la racine n'est pas trouvée, au lieu de traiter une chaîne de caractères de longueur $R-1$ (ou $R+1$ pour une recherche progressive), elle prend en compte $R \pm S_i$ caractères comme indiqué ci-dessous :

1) $\sum_{i=1}^l R = R + S_i$ (recherche progressive, R : racine ; S_i : i -ème syllabe ; i : nombre de syllabe ; l : dernier syllabe)

2) $\sum_{i=1}^l R = R - S_i$ (recherche régressive, R : racine ; S_i : i -ème syllabe ; i : nombre de syllabe ; l : dernier syllabe)

Les règles de syllabisation sont assez complexes (voir chapitre 6 pour plus de détails). Sur la base de plusieurs tests, nous avons mis au point de façon empirique un algorithme de syllabisation dont nous estimons qu'il donne les résultats les plus rapides et les plus fiables dans son exécution. Pour qu'il soit prêt à tester pour les amateurs de traitement automatique de la langue ouïghoure, nous le présentons également en code C# :

Début

recupérer les voyelles de l'entrée et la chaîne de caractères à syllabiser

initialisation à vide des variables: positions de la 1^{ère} et 2^{ème}

voyelle, longueur de la chaîne de caractères à syllabiser, première syllabe, reste des syllabes, 1^{ère} et 2^{ème} voyelle, stockage du résultat

SI la longueur des voyelles de l'entrée est inférieure ou égale à un ALORS

retourner la chaîne de caractères à syllabiser telle quelle : cela indique qu'il n'y a qu'une seule

ou pas de voyelle

SINON

recupération de la première voyelle

recupération de la seconde voyelle

recupération de position de la première voyelle

recupération de position de la seconde voyelle

SI la position de la 1^{ère} voyelle est égale à la position de la lettre qui précède la 2^{ème}

----- ALORS

Algorithme 3. Syllabisation des mots ouïghours

```

// Uyghur Syllablizer (c) 2007 Waris A. Janbaz, oyghan@gmail.com
// s: la chaîne de caractères à syllabiser ; v : tous les voyelles de « s » ;
public string UyghurSyllablizer(string s, string v)
{
    // initialisations
    int l = s.Length, idx1 = 0, idx2 = 0;
    string fv, sv, slb = "", lstr, rstr;

    if (v.Length <= 1) // s'il n'y a qu'une seule ou pas de voyelle dans « s »
    {
        return s;
    }
    else
    {
        fv = v.Substring(0, 1); // première voyelle
        sv = v.Substring(1, 1); // seconde voyelle
        idx1 = s.IndexOf(fv, idx1); // position de la première voyelle
        idx2 = s.IndexOf(sv, idx1 + 1); // position de la deuxième voyelle
        if (idx1 == idx2 - 1) idx2++; // éviter des syllabes non ouïghoures
        lstr = s.Substring(0, idx2 - 1); // première syllabe de « s »
        // reste des syllabes après soustraction de la première syllabe de « s »
    }
}

```

Code source en C# de l'algorithme 3.

La syllabisation est un procédé inévitable pour les logiciels de synthèse vocale. Avec l'approche syllabique, la mise au point d'un « outil générique » de synthèse de la parole à partir du texte (*text-to-speech*) est envisageable car il y a beaucoup plus mots que des syllabes. Pour cela, il faut créer une base de représentation phonétique de chaque syllabe ouïghoure, et « réécrire » chaque syllabe du texte original en utilisant sa représentation dans la base de données vocales. Cependant, comme la synthèse vocale ne fait pas l'objet de notre étude, nous ne la mentionnons ici qu'à titre de référence.

8.5 Vérification d'orthographe

Un mot est correctement écrit (épelé) si 1) il se trouve dans le dictionnaire de base ; 2) il est dans la liste des racines primitives ; ou 3) sa racine se trouve dans la liste des racines

primitives et sa partie suffixe est dans la liste des suffixes. Nous avons étudié dans les paragraphes précédents la première partie du troisième cas. Une fois la racine déterminée, il faut vérifier la partie suffixe pour voir si elle se conforme aux règles de suffixation afin d'être sûr que l'orthographe du mot entier est correcte. Notre analyseur lexical (voir 5.9.4) permet non seulement de vérifier la présence de fautes d'orthographe, mais aussi d'indiquer leur nature exacte, à savoir la graphie fautive et le numéro de la syllabe dans laquelle elle apparaît. Dans notre outil prototype, prenant en considération le fait que une grande majorité des utilisateurs ne s'intéresse pas à savoir dans laquelle partie de la chaîne de suffixes l'erreur est survenue, nous avons traité la partie suffixe comme une entité, ce qui nous permet de ne pas ralentir la procédure de traitement. Nous avons construit, à l'aide du générateur de lexique (voir 5.9.3), huit listes²²⁴ typologiques de suffixes qui couvrent la quasi-totalité des combinaisons possible. Lors de la vérification, la partie suffixe du mot sera recherchée dans une des huit listes selon les caractéristiques phonétiques de la racine. Par exemple, pour le mot «*asmanglar*» (n'accrochez pas) une fois sa racine «*as*» identifiée, sa partie suffixe – «*manglar*» ne sera recherchée que dans la liste BUH (suffixes des racines qui contiennent des voyelles arrières non-arrondies et terminées par une consonne dure). La figure 8.4 ci-dessous donne une vue schématique de l'ensemble de la procédure de vérification d'orthographe.

²²⁴ A savoir: FRH, FRS, FUH, FUS, BRH, BRS, BUH, BUS. F: voyelle d'avant; B: voyelle d'arrière ; R : voyelle arrondie ; U : voyelle non-arrondie ; H : consonne dur ; S : consonne douce. Voir 5.2 consonnes et 5.3 voyelles du chapitre 5.

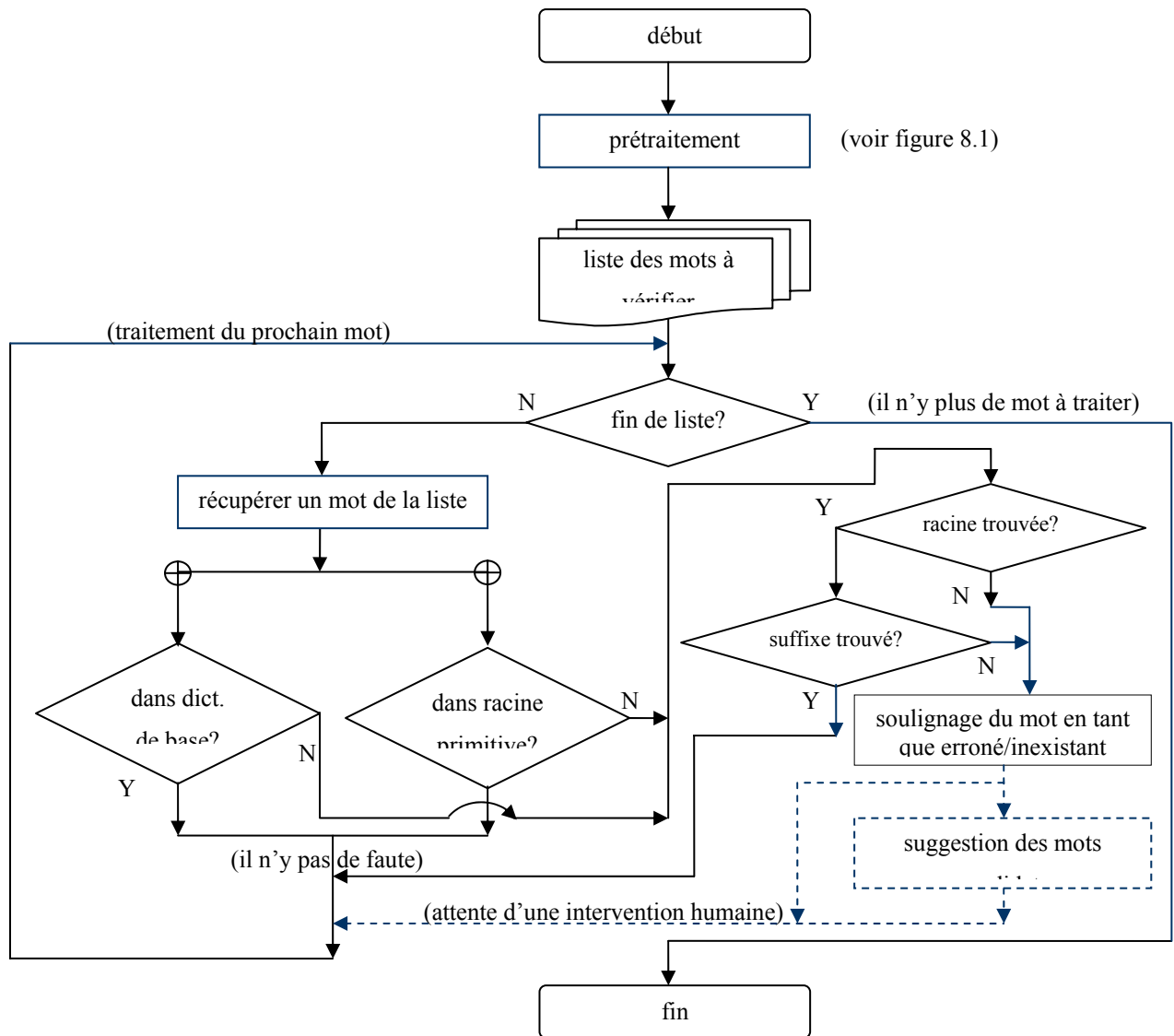


Figure 8.4 vérification d'orthographe

Dans cette figure, les lignes pointillées indiquent que cette étape est facultative et son application exige une intervention humaine pour sélectionner un des mots suggérés ou corriger manuellement. L'étape de suggestion des mots candidats pour des mots erronés est expliquée dans le chapitre 4.9.5.

Chapitre 9. Conclusion et travaux futures

Dans cette étude, notre objectif a été de présenter aux lecteurs les caractéristiques de la langue et des écritures ouïghoures et d'offrir aux utilisateurs des outils informatiques qui permettent de contribuer à l'informatisation et à la sauvegarde de cette langue menacée – l'ouïghour. Au cours de la première année de notre étude nous nous étions orientés vers le développement d'un moteur de recherche trans-langue (ouïghour-anglais) et multiécriture. C'est cette orientation originale qui est à l'origine des études préparatoires figurant dans les chapitres 2 et 7. La visualisation du résultat d'analyse et d'extraction des entités nommées a été réalisée sur l'Unitex avec utilisation d'automates et de transducteurs. Ces études, effectuées dans une phase expérimentale, nous ont fait comprendre l'importance et la nécessité d'une analyse linguistique pour la mise en place des outils de traitement automatique, moteur de recherche compris, pour l'ouïghour. C'est dans ce but que nous avons approfondi nos recherches sur une approche linguistique visant à établir une méthodologie d'extraction d'information appliquée à l'analyse linguistique.

Dans le chapitre 6 nous avons fait l'analyse de la morphologie de l'ouïghour, en tenant compte de problématiques comme l'harmonie vocalique et l'affaiblissement de voyelles dans leurs rapports avec le traitement automatique de la langue. Compte tenu de la complexité des règles de suffixation, nous avons consacré le chapitre 5 à une description détaillée de ce phénomène fondamental. Tous les efforts que nous avons faits n'ont pas, comme nous l'aurions souhaité dans un premier temps, abouti à l'élaboration de produits concrets mais ils nous ont permis de débayer un certain nombre d'obstacles et d'ouvrir ainsi des pistes fiables pour des recherches pratiques ultérieures destinées à la réalisation de ces produits. C'est donc avec un mélange d'humilité mais aussi d'espoir face aux tâches qui restent à accomplir que nous résumons ci-dessous le nouvel état des lieux du sujet à l'issue de nos recherches. :

- **Ecriture ouïghoure-latine**

Au cours des dernières années, le LSU (alphabet latin de d'ouïghour) a été adopté pour des raisons pratiques dans les échanges informatiques. Les lettres des trois systèmes alphabétiques en usage pour l'ouïghour (arabe, cyrillique et latin) se correspondent entre elles et les règles d'orthographe, dans l'ensemble, ne varient pas de l'un à l'autre.

Indifféremment du succès du LSU, certains restent prudents à son égard par peur qu'il ne finisse par se substituer à l'ASU, basé sur l'Arabe et héritier d'une tradition multiséculaire. Ils estiment que le progrès dans le traitement informatique de systèmes d'écriture non latines, particulièrement grâce à l'avènement de l'Unicode, a éliminé le besoin d'une transcription latine.

Au vingtième siècle, les Ouïghours ont eu plus que leur part de changements d'alphabets, ce qui a entraîné une coupure d'avec leur héritage culturel et la création de nouveaux obstacles à l'acquisition de connaissances. La multiplication artificielle des systèmes d'écriture a de plus creusé les frontières géopolitiques et le fossé intergénérationnel.

En tant que système de transcription simple, logique et clair, le LSU a renforcé les activités d'échange et le traitement d'information en langue ouïghoure. Il donne ainsi la chance aux Ouïghours de rattraper le monde dans leur propre langue au lieu d'avoir à passer par l'intermédiaire d'une autre. Le progrès du traitement d'information et des échanges en ASU ne peut que bénéficier des avances faites grâce à LSU.

Plus d'efforts sont nécessaires afin de développer des instruments adéquats pour le traitement de l'information en langue ouïghoure. Les fautes d'orthographe dues à une mauvaise saisie du LSU, ainsi que les règles d'orthographe en ASU, sont des problèmes importants pour l'implémentation de correcteurs d'orthographe et d'outils de conversion des écritures. Comme le bas taux d'alphabétisation et la prolifération de réformes orthographiques ont eu pour conséquence une situation chaotique au niveau de l'orthographe dans la population, la révision des programmes de conversion inter-alphabet et la création d'un outil complet de correction orthographique sont de tâches de toute première priorité.

Le LSU s'est créé une niche spéciale dans la communauté internationale. Une masse critique se forme autour de lui. Il n'a pas remplacé l'ASU mais il joue un rôle complémentaire par rapport à ce dernier. Il a permis à la langue ouïghoure d'être présente sur le web et dans les publications de manière lisible. Espérons qu'il remplacera les vieilles translittérations et les versions sinisées des noms propres et géographiques ouïghours dans la presse et la littérature internationales..

- **Traitement de la langue ouïghoure**

Notre travail dans ce domaine a surtout eu pour but le design et la résolution des problèmes de mise en œuvre dans le cadre de la création des polices ouïghoures, ainsi que les applications de conversion multi-écriture et les méthodes d'entrées à deux niveaux (au niveau du système d'exploitation et au niveau de l'explorateur). Les commentaires des utilisateurs semblent indiquer que cette première recherche sur le traitement de l'ouïghour a eu des résultats satisfaisants. Les polices intégrées pour le web, générées par un logiciel tiers WEFT, sont uniquement compatibles avec Internet Explorer. Nous attendons donc avec une certaine impatience que l'industrie informatique fasse les efforts nécessaires à une plus grande compatibilité. Nous espérons pouvoir améliorer le module de prétraitement de l'outil de conversion afin de le rendre plus « convivial ».

Nous avons développé un dictionnaire en ligne afin mettre en valeur les résultats de nos études sur l'implémentation des solutions pour du traitement de la langue ouïghoure sur le web. Les commentaires des utilisateurs du dictionnaire et le nombre des visiteurs du site sont très encourageants. Cependant, un problème relié à la suggestion des mots-clés (lorsque le mot clé saisi par l'utilisateur n'a pas été retrouvé) est le fait qu'il s'agit d'une langue agglutinante, un obstacle majeur dans le développement d'une fonctionnalité de correcteur orthographique et qui est rendu plus complexe par le foisonnement de suffixes le phénomène, unique à l'ouïghour, des changements orthographiques pouvant affecter les racines de mots en présence de suffixes. Nous estimons pouvoir balayer les obstacles avec l'application des règles et des méthodes présentées dans le chapitre 5 et 8. Résoudre ce problème sera le but de notre travail dans la prochaine étape du développement.

Nous avons insisté sur l'historique du développement des logiciels en ouïghour et des questions de mise en œuvre liées aux polices ouïghoures Unicode et IME. Les efforts investis, au cours de plus de trois ans, dans l'unification des polices ouïghoures en utilisant le standard Unicode ont été fructueux. Beaucoup d'applications et de sites web ont été créés afin de développer la technologie d'information ouïghoure. Il serait nécessaire que les autorités gouvernementales de la Région Autonome soutiennent plus efficacement ce travail en appliquant véritablement la loi de langue et d'écriture de la RAOX, afin de satisfaire la demande croissante de logiciels et sites web en ouïghour. En même temps, les compagnies informatiques devraient davantage s'efforcer d'offrir des produits plus compatibles. Il y a aussi d'autres problèmes techniques à résoudre afin de pouvoir ajouter des définitions de

locales ouïghoures à Windows XP, ainsi qu'aux versions plus anciennes. Nous faisons appel aux compagnies de logiciels pour qu'elles n'omettent plus la langue ouïghoure dans la liste des langues supportées.

- **Règles morphologiques de suffixation**

Nous avons présenté une vue d'ensemble des règles de suffixation et d'un analyseur morphologique du ouïghour en utilisant des verbes en tant qu'échantillon représentatif. Cette approche n'a jamais été utilisée et a été jusqu'à présent ignorée dans le traitement de la langue ouïghoure. Nous croyons que cette méthodologie peut être appliquée à un logiciel TAL, particulièrement : 1) dans les moteurs de recherche lors de l'indexation et l'analyse des requêtes de recherche ; 2) dans les correcteurs orthographiques ou logiciels OCR afin de vérifier/suggérer les orthographes correctes/possibles ; 3) dans les méthodes d'entrée afin d'accélérer la rapidité d'entrée ; 4) dans les dictionnaires afin de réduire les entrées de surfaces répétitives. Comme nous pouvons le constater dans le chapitre 5, en utilisant cette méthode linguistique nous pouvons obtenir jusqu'à 24,294 formes dérivées différentes pour une seule racine verbale. En conséquence de quoi, cette approche aide à réduire le nombre de formes dérivées de verbes dans un corpus. Puisque la banque de règles couvre toutes les suites de suffixes pour tous les verbes, certaines des formes verbales obtenues sont théoriquement possibles mais ne sont pas utilisées.

Même si le générateur lexical est capable de créer de nouveaux mots et que l'explorateur de suffixes peut expliquer les modifications de radical, les frontières de suffixes et l'ordre de suffixation, ils ne constituent pas encore un logiciel directement utilisable. Des travaux sur les racines nominales et les règles de dérivation doivent encore être effectués avant de pouvoir parvenir à la mise au point de logiciels pour l'ensemble du lexique ouïghour.

- **Parseur et correcteur d'orthographe**

Après une brève analyse des correcteurs d'orthographe ouïghours existants, nous avons expliqué pourquoi une nouvelle approche était nécessaire pour couvrir tous les mots théoriquement possibles dans des outils de traitement automatique de la langue ouïghoure, plus particulièrement, dans un outil de correcteur d'orthographe. Ensuite, nous avons étudié le fonctionnement d'un parseur et des opérations qu'il effectue sur les objets de traitement

pendant le processus de prétraitement. Nous avons adapté les deux types de recherche classique – progressive et régressive – à l’identification de la racine. Nous avons réussi à réutiliser les règles de suffixation définies pour notre générateur de lexique (voir chapitre 5.9.3) pendant l’étape de dé-suffixation afin de trouver facilement la racine des mots hautement suffixés. Nous avons créé un algorithme qui permet de syllabiser des mots ouïghours selon les règles lexicales. Enfin, dans une phase d’expérimentation, les méthodes mentionnées ont été implémentées dans un outil-prototype de vérification d’orthographe des dérivés verbaux à des fins de vérification pratique. L’outil complet n’a pas encore pu être mis en place : il s’agissait pour nous d’établir une méthodologie satisfaisante pour la catégorie lexicale la plus complexe de la langue ouïghoure, les verbes, en raison de la multiplicité des possibilités de dérivation qu’ils offrent. L’application de cette méthodologie aux autres catégories lexicales sujettes à dérivation et suffixation, à savoir les noms et les adjectifs, ne devrait pas poser de problème particulier, puisque les mêmes principes s’appliquent dans leur cas mais avec un nombre de suffixes grandement réduit. Comme nous pouvons le constater, la stratégie de recherche pour développer un outil complet de correcteur orthographique a déjà été mise en place. Nous espérons pouvoir les développer dans des recherches futures.

Chapitre 10. Bibliographie

- [1] Dr. Rıdvan ÖZTÜRK, Yeni Uygur Türkçesi Gramerı, ANKARA, Semih Ofset Matbaacılık ve Ambalaj Sayayı Ltd., 1994, ISBN: 975-16-0586-5. (en turc)
- [2] Dr. Sultan Mehmet Kaşgarlı, Modern Uygur Türkçesi Gramerı, Istanbul, 1992, Kardeşler Matbaası, ISBN: 975-7740-12-8 (en turc)
- [3] Qazaq SSR Penler Akadémiyisi Uyghurshunasliq Bölimi, Hazırqi Zaman Uyghur Tili, 1-qisim, Léksika we Fonétika, Alma-ata, édition 1966, Shinjang Xelq Neshiryati, 1985, M9098.40 (en ouïghour)
- [4] Qazaq SSR Penler Akadémiyisi Uyghurshunasliq Bölimi, Hazırqi Zaman Uyghur Tili, 2-qisim, Morfologiyе we Sintaksis, Alma-ata, 1966, Shinjang Xelq Neshiryati, 1985, M9098.37 (en ouïghour)
- [5] Alfred Morer, Grammaire de la Langue Turque (théorique et pratique), 8^{ème} édition, 1986
- [6] Office québécois de la langue française, Revue d'Aménagement Linguistique, Numéro 106, été 2003, Saint-Laurent – Canada, ISSN 1706-3515.
- [7] Serge Viallet, Oasis perdues des Routes de la Soie (CD-ROM), Paris, 1995, Réunion des Musées Nationaux, code bar : 3-336725-000122
- [8] Jean-Paul Roux, Histoire des Turcs, 1991, France, Fayard, ISBN : 2-213-01491-4, 35-65-7268-0512
- [9] Haji Nurhaji, Qedimki Uyghurlar we Qaraxaniylar, Ürümchi, 2001, Shinjang Xelq Neshiryati, ISBN : 7-228-06389-9/K.927 (en ouïghour).
- [10] Reinhard F. Hahn, Spoken Uyghur, Washington, 1991, the University of Washington Press, ISBN: 0-295-97015-4.
- [11] Abdushükür Muhemmetimin, Qedimki Merkiziy Asiya, Ürümchi, 2002, Shinjang Xelq Neshiryati, ISBN: 7-228-06737-1 (en ouïghour).
- [12] Rehmetjan Imin, Uyghur tili, Ürümchi, 2001, 1-qisim, Shinjang Xelq Neshiryati, ISBN: 7-5370-3521-0 (en ouïghour).
- [13] Abdulétip Tashpolat, Uyghur tili, Ürümchi, 2001, 2-qisim, Shinjang Xelq Neshiryati, ISBN: 7-5370-3520-2 (en ouïghour).

- [14]Abdulétip Tashpolat, Uyghur tili, Ürümchi, 2001, 3-qisim, Shinjang Xelq Neshiryati, ISBN: 7-5370-3517-2 (en ouïghour).
- [15]Christian Baylon & Paul Fabre, Initiation à la linguistique, Paris, 1990, Édition Fernard Nathan, ISBN : 2-09-190765-9.
- [16]Sh.U.A.R Tilkom, Hazirqi zaman Uyghur edemiy tilining imla lughiti, Ürümchi, 1985, Shinjang Xelq Neshiryati, M9098.42 (en ouïghour).
- [17]Xemit Tömür, Hazirqi zaman Uyghur tili grammatikisi (morphologiye), Béyjing, 1987, Milletler Neshiryati, M9049 (4)11 (en ouïghour)
- [18]Sh.U.A.R Tilkom, Uyghur tilining izahliq lughiti, Ürümchi, 1999, Shinjang Xelq Neshiryati, ISBN : 7-228-05448-2/H.146 (en ouïghour).
- [19]Qurban Weli, Bizning Tarixiy Yéziqlirimiz,Ürümchi, Xinjiang Youth's Press, 1986, (en ouïghour).
- [20]Amine Ghappar, Mirsultan Osmanov etc., Hazirqi zaman Uyghur adebiy tilining teleppuz lughiti, Ürümchi, 1988, Milletler neshiryati, ISBN : 7-105-00529-7/h.28, (en ouïghour).
- [21]Kurtuluş Öztopçu & Zhoumagaly Abouov & Nasir Kambarov & Youssef Azemoun, Dictionary of the Turkic Languages (Azerbaijani, Kazakh, Kirghiz, Tatar, Turkish, Turkmen, Uyghur, Uzbek), New York, reprinted 1999, Routledge, ISBN: 0-415-14198-2 (hbk), ISBN: 0-415-16047-2 (pbk).
- [22]Zhao Xiang Ru & Zhu Zhi Ning, wei wu er yu jian zhi, Pékin, 1985, Min zu chu ban she, BN: 9049.40 (en chinois).
- [23]Geng Shi Min & Li Zeng Xiang, ha sa ke yu jian zhi, Pékin, 1985, Min zu chu ban she, BN: 9049.51 (en chinois).
- [24]Dao Bu, meng gu yu jian zhi, Pékin, 1985, Min zu chu ban she, BN: 9049.29 (en chinois)
- [25]Chen Zong Zhen & Yi Li Qian, ta ta er yu jian zhi, Pékin, 1985, Min zu chu ban she, BN: 9049.75 (en chinois).
- [26]Chen Shi Liang & Abdurahman, wu zi bie ke yu jian zhi, Pékin, 1985, Min zu chu ban she, BN: 9049.77 (en chinois).
- [27]Hu Zhen Hua, ke er ke zi jian zhi, Pékin, 1985, Min zu chu ban she, BN: 9049.64 (en chinois).
- [28]Richard Sproat, Morphology and Computation, USA, 1992, MIT press, ISBN: 0-262-19314-0

- [29] AFNOR, Principes généraux pour l'indexation des documents, tome 1, Paris la Défense, 1996, AFNOR NF Z 47-102
- [30] Suzanne Walter, L'analyse documentaire, Paris, 1999, ADBS, ISBN : 2-84365-030-5
- [31] Chen. Multilingual Information Retrieval Using English and Chinese Queries, Proceedings of CLEF-2001, Sophia-Antipolis: ERCIM EEIG, 2001.
- [32] Gérard Sabah, L'intelligence artificielle et le langage, volume 1, Paris, 1988, Hermès, ISBN : 2-86601-134-1
- [33] Frédérique Segond etc., Multilinguisme et traitement de l'information, Paris, 2002, LAVOISIER, ISBN : 2-7462-0523-8
- [34] François Rastier & Marc Cavazza & Anne Abeillé, Sémantique pour l'analyse de la linguistique à l'informatique, Paris, 1994, MASSON, ISBN : 2-225-84537-9
- [35] Thierry Poibeau, Extraction automatique d'information, Paris, 2003, Hermès, ISBN 2-7462-0610-2.
- [36] Thierry Poibeau, «Extraction d'information dans les bases de données textuelles en génomique au moyen de transducteurs à nombre fini d'états », Conférence Française de Traitement Automatique de la Langue, (TALN'2001),
- [37] http://www.li.univ-tours.fr/taln-recital-2001/Actes/tome1_PDF/partie2_p30_322/art27_p293_302.pdf.
- [38] Christine Michel, «Evaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogènes », thèse doctorat de l'Université Lumière Lyon II, 1999, <http://www.recodoc.univ-lyon1.fr/theseCMichel.pdf>.
- [39] Marie Calberg, « Traitement de la morphologie du finnois par transducteur à nombre fini d'états », Batz-sur-Mer, RÉCITAL 2003, http://www.cavi.univ-paris3.fr/ilpga/ED/activites/RJC2003_actes/calberg.pdf
- [40] Christof Monz & Maarten de Rijke , Introduction to Information Retrieval, Amsterdam , 2002 (une présentation pour l'enseignement, 231 pages) <http://remote.science.uva.nl/~christof/courses/ir/transparencies/w-01-prst.pdf>.
- [41] Kemal Oflazer & Gökhan Tür & Dilek Hakkani Tür, A statistical information extraction system for Turkish, Ankara, 2001, <http://www.research.att.com/~dtur/pubs/NLE03.pdf> .

- [42] Kemal Oflazer & İlker Kuruöz, A Tool for Tagging Turkish Text, Ankara, 1994, <http://www.cs.bilkent.edu.tr/tech-reports/1994/BU-CEIS-9416.ps.z>
- [43] Gökhan Tür & D. Z. Hakkani-Tür & Kemal Oflazer, Statistical Modeling of Turkish for Automatic Topic Segmentation, Ankara, 2000, <http://www.cs.bilkent.edu.tr/tech-reports/2000/BU-CE-0001.ps.gz>.
- [44] Sébastien Paumier, Unitex Manuel d'Utilisation, Paris, 2002, <http://www-igm.univ-mlv.fr/~unitex/>
- [45] Kemal Altintas & Ilyas Cicekli, A Morphological Analyser for Crimean Tatar, http://www.ics.uci.edu/~kemal/publications/ct_morphology_altintas_cicekli.ps
- [46] Claude de Loupy & Patrice Bellot, Evaluation of Document Retrieval Systems and Query Difficulty, Avignon, http://www.lia.univ-avignon.fr/fich_art/146-lreceval.pdf
- [47] Claude de Loupy, L'apport de connaissances linguistiques en recherche documentaire, Ivry-sur-Seine, TALN 2001, <http://www.sinequa.com/sq-lab-doc/SL-2001-007-TALN2001.pdf>
- [48] Alim Ahat, Frequency analyses of the Uyghur letters, Jan 2004. <http://www.uyghursoft.com/uyghur/soft/biliwal.htm>
- [49] Waris A. Janbaz, *Kompyutér saheside "Latin yéziqi asasadiki Uyghur yéziqi" ni tesis qilish toghrisida teshebbusname* (29-page document distributed during the first conference), Nov 2000.
- [50] Uyghur Computer Science Association, "*Uyghur kompyutér yéziqi toghrisida*", http://www.ukij.org/teshwiq/UKY_Heqqide.htm, May 2004.
- [51] UKY *Arxipliri* (discussion archives on internet 200-2001), www.biliwal.com/modules.php?name=UKY_Arxipi (also available at: www.voy.com/37018).
- [52] Website of the XUAR Working Committee of Minorities' Language and Writing, wénzìgǎigé (writing reforms), http://www.xjyw.gov.cn/han/wenzi_gai.htm, 2001.
- [53] J. R. Duval: "Modern Uyghur, A Historical Perspective." In: Culture Contact, History and Ethnicity in Inner Asia, № 2. Michael Gervers and Wayne Schlepp, eds. Toronto: Joint Center for Asia Pacific Studies, 1996, pp. 132-67.
- [54] Jean Rahman Duval, Waris Abdukérím Janbaz, September 2006, "An Introduction to Latin-Script Uyghur", 2006 Middle East & Central Asia Conference, University of Utah, Salt Lake City, USA.

- [55] Jean Rahman Duval, Waris Abdukérím Janbaz, July 2005, “The Uyghur English *Dictionary*”, <http://www.uyghurdictionary.org>
- [56] Jean Rahman Duval, 1996, “*Modern Uyghur: A Historical Perspective*”, in *Cultural Contact, History and Ethnicity in Inner Asia*, edited by Michael Gevers and Wayne Schlepp, Joint Centre for Asia Pacific Studies, Toronto, ISBN 1-895296-22-6
- [57] Reinhard F. Hahn, 1991, *Spoken Uyghur*, University of Washington, ISBN: 0-295-97015-4.
- [58] Ayşin Solak, Kemal Oflazer, 1993, “*Design and Implementation of a spelling checker for Turkish*”, *Literary and linguistic computing*, Vol. 8, No. 3, Oxford University Press.
- [59] Kemal Oflazer, 1994, *Two-level Description of Turkish Morphology*, *Literary and Linguistic Computing*, Vol. 9, No:2.
- [60] Dr. Rıdvan ÖZTÜRK, 1994, *Yeni Uyğur Türkçesi Gramerı*, ANKARA, Semih Ofset Matbaacılık ve Ambalaj Sayayı Ltd., ISBN: 975-16-0586-5 (in Turkish).
- [61] Dr. Sultan Mehmut Kaşgarlı, *Modern Uyğur Türkçesi Gramerı*, İstanbul, 1992, Kardeşler Matbaası, ISBN: 975-7740-12-8 (in Turkish).
- [62] *Qazaq SSR Penler Akademiyisi Uyğurshunasliq Bölimi*, Alma-ata, 1966, *Hazirqi Zaman Uyğur Tili, 1-qisim, Léksika we Fonétika*, Shinjang Xelq Neshiryati, 1985, M9098.40 (in Uyghur).
- [63] *Qazaq SSR Penler Akademiyisi Uyğurshunasliq Bölimi*, Alma-ata, 1966, *Hazirqi Zaman Uyğur Tili, 2-qisim, Morfologiyeye we Sintaksis*, Shinjang Xelq Neshiryati, 1985, M9098.37 (in Uyghur).
- [64] Sh.U.A.R Tilkom, 1985, *Hazirqi zaman Uyğur edebiy tilining imla lughiti*, Ürümchi, Shinjang Xelq Neshiryati, M9098.42 (in Uyghur).
- [65] Xemit Tömür, 1987, *Hazirqi zaman Uyğur tili grammatikisi (morfologiyeye)*, Beijing, Milletler Neshiryati, M9049 (4)11 (in Uyghur).
- [66] Sh.U.A.R Tilkom, 1999, *Uyğur tilining izahliq lughiti*, Ürümchi, Shinjang Xelq Neshiryati, ISBN : 7-228-05448-2/H.146 (in Uyghur).
- [67] Amine Ghappar, Mirsultan Osmanov etc., *Hazirqi zaman Uyğur adebiy tilining teleppuz lughiti*, 1988, Ürümchi, Milletler neshiryati, ISBN : 7-105-00529-7/h.28 (in Uyghur).
- [68] Richard Sproat, 1992, *Morphology and Computation*, USA, MIT press, ISBN: 0-262-19314-0.

- [69] Cüneyd Tantuğ and Esref Adalı and Kemal Oflazer, 2006, *Computer Analysis of the Turkmen Language Morphology*, FinTAL, *Lecture Notes in Computer Science*, Vol. 4139, pp. 186-193, Springer.
- [70] Waris A. Janbaz, *Online Uyghur Unicode processing technique and its implementation* (publication in Chinese), Xinjiang University Press, China, 2002.
- [71] Abdurehim, Waris A. Janbaz, *Orthographic rules of the Latin-Script Uyghur (in Uyghur)*, 2004, [http://www.ukij.org/teshwiq/UKY_Hegqide\(KonaYeziq\).htm](http://www.ukij.org/teshwiq/UKY_Hegqide(KonaYeziq).htm).
- [72] The Unicode Consortium *The Unicode Standard, Version 4.0*, Addison-Wesley Professional, ISBN: 0321185781, USA, 2003.
- [73] Xinjiang University, *Proceedings 2000 International Conference on Multilingual Information Processing*. Ürümchi (publication in Chinese), China, 2000.
- [74] The Unicode Consortium Website <http://www.unicode.org>
- [75] Reinhard F. Hahn, *Spoken Uyghur*. Washington: the University of Washington Press, ISBN: 0-295-97015-4, USA, 1991.
- [76] Mohamed Hatem HADDAD, *Extraction et Impact des connaissances sur les performances des Systèmes de Recherche d'Information*, 2002, <http://hal.archives-ouvertes.fr/docs/00/04/60/54/PDF/tel-00004459.pdf>
- [77] Claude de Loupy, *L'apport de connaissances linguistiques en recherchedocumentaire*, 2001, <http://www.syllabs.com/perso/loupy/cv/papers/LOU01a.pdf>
- [78] Christine Michel, *Evaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogènes*, thèse 1999, <http://www.recodoc.univ-lyon1.fr/theseCMichel.pdf>

Chapitre 11. Annexes

Annexe 1. Carte des groupes ethniques de l'Asie Centrale

Annexe 2. Alphabet turco - runique (VIème - IXème)

Annexe 3. Alphabet ouïghour ancien (Xème-XVIIIème)

Annexe 4. Table des alphabets de langues turciques

Annexe 5. Abréviations

Annexe 6. Suffixes ouïghours

Annexe 7. Lettres ouïghoures et leurs valeurs Unicode conventionnées

Annexe 8. Groups des suffixes verbaux

Annexe 9. Liste des verbes primitive

Annexe 10. Tableau des alphabets ASU, LSU, CSU

Annexe 11. Arbre des langues turciques

Chapitre 11. Annexes

Tableau des alphabets ASU, LSU, CSU

ف	غ	ش	س	ژ	ز	ر	د	خ	چ	ج	ت	پ	ب	ئ	ئا	ASU
f	gh	sh	s	j /zh	z	r	d	x	ch	j	t	p	b	e	a	LSU
ф	Ғ	Ш	с	Ж	з	р	д	х	ч	ж	т	п	б	е	а	CSU
ي	ئى	ئې	ۋ	ئۇ	ئۆ	ئو	ھ	ن	م	ل	ڭ	گ	ك	ق		ASU
y	i	é	w	ü	ö	u	o	h	n	m	l	ng	g	k	q	LSU
й	и	е	в	ү	ө	у	о	һ	н	м	л	ң	г	к	қ	CSU

Lettres cyrilliques additionnelles : ы ё ц э ю я

ئۇيغۇر تىلى ئېلىپبەسى

 پ پ پ پ پ	 ب ب ب ب ب	 ئە ئە ئە ئە	 ئا ئا ئا ئا
 خ خ خ خ خ	 چ چ چ چ چ	 ج ج ج ج ج	 ت ت ت ت ت
 ژ ژ ژ ژ ژ	 ز ز ز ز ز	 ر ر ر ر ر	 د د د د د
 ف ف ف ف ف	 غ غ غ غ غ	 ش ش ش ش ش	 س س س س س
 ك ك ك ك ك	 گ گ گ گ گ	 ك ك ك ك ك	 ق ق ق ق ق
 ھ ھ ھ ھ ھ	 ن ن ن ن ن	 م م م م م	 ل ل ل ل ل
 ئو ئو ئو ئو ئو	 ئو ئو ئو ئو ئو	 ئو ئو ئو ئو ئو	 ئو ئو ئو ئو ئو
 ي ي ي ي ي	 ئى ئى ئى ئى ئى ئى	 ئى ئى ئى ئى ئى ئى	 قۇ قۇ قۇ قۇ قۇ

Carte des groupes ethniques de l'Asie Centrale

Alphabet turco-runic (VI^{ème} - IX^{ème})

Alphabet ouïghour ancien (X^{ème}-XVIII^{ème})

noms des suffixes ouïghours

K=[R]i	euphonic	BR=ghach
L=[R]i	AO=i	BS=qach
M=[R]ü	AP=i	aspectual auxiliary
N=[R]u	AQ=ü	BT=wat
reflexive	AR=u	negative
O=n	2nd collective	BU=ma
P=n	AS=sh	interrogative
passive	AT=sh	BV=em
Q=l	Eupho. Con.	BW=m
R=l	AU=y	BX=am
Euphonic-	AV=y	BY=m
S=i	potential	Tense: infinitive
T=i	AW=ele	BZ=mek
U=ü	AX=eli	CA=mik
V=u	AY=el	CB=maq
collective	AZ=ala	CC=miq
W=sh	BA=ali	declinable
X=sh	BB=al	infinitive
factitive	negative	(substantiative)
Y=dür	BC=me	CD=sh
Z=tür	BD=mi	CE=sh
AA=dur	BE=ma	desiderative
AB=tur	BF=mi	CF=gü
euphonic	repetitive	CG=kü
AC=i	BG=e	CH=ghu
AD=i	BH=a	CI=qu
AE=ü	connective	participial past
AF=u	BI=p	CJ=gen
passive	BJ=y	CK=ken
AG=l	BK=w	CL=ghan
AH=l	purpositive	CM=qan
aspectual auxiliary	BL=gili	CN=gin
AI=wet	BM=kili	CO=kin
AJ=wét	BN=ghili	CP=ghin
AK=wer	BO=qili	CQ=qin
AL=wér	simultaneative	adverbialive
AM=wal	BP=gech	CR=séri
AN=wél	BQ=kech	

desiderative

CS=gey

CT=key

CU=ghay

CV=qay

polite imperative

CW=gin

CX=kin

CY=ghin

CZ=qin

imperative

DA=sun

DB=ey

DC=y

DD=ay

DE=y

habitual-past

DF=et

DG=yt

DH=at

DI=yt

future

DJ=i

DK=y

future participle

DL=dighan

DM=dighin

positive**speculative future**

DN=er

DO=r

DP=ar

DQ=r

negative**speculative future**

DR=s

DS=s

conditional

DT=se

DU=si

DV=sa

DW=si

deverbal agent**noun/adjective****former**

DX=chi

DY=chi

DZ=ch

EA=ch

privative

EB=siz

substantive**(adjective former)**

EC=lik

ED=lük

EE=liq

EF=luq

pre-past tense**auxiliary****interrogative****particle**

EG=m-

neutral auxiliary**used for past tense****formation**

EH=i

perfective past

EI=di

EJ=ti

EK=dü

EL=tü

EM=du

EN=tu

euphonic

EO=i

EP=i

EQ=ü

ER=u

personal possessive

ES=m

ET=m

EU=ng

EV=ng

EW=ngiz

EX=ngiz

EY=i

EZ=si

FA=i

FB=si

FC=miz

FD=miz

FE=k

FF=q

FG=nglar

FH=liri

FI=liri

relative locative

FJ=diki

FK=tiki

FL=diki

FM=tiki

plural

FN=ler

FO=lir

FP=lar

FQ=lir

person**possessive (after****plural)**

FR=im

FS=im

FT=ing

FU=ing

FV=ingiz

FW=ingiz

FX=i

FY=i

FZ=imiz

GA=imiz
 GB=inglar
locative
 GC=de
 GD=di
 GE=te
 GF=ti
 GG=da
 GH=di
 GI=ta
 GJ=ti
dative
 GK=ge
 GL=gi
 GM=ke
 GN=ki
 GO=gha
 GP=ghi
 GQ=qa
 GR=qj
ablative
 GS=din
 GT=tin
accusative
 GU=ni
genitive
 GV=ning
pronominal
genitive
 GW=ningki
declarative
 GX=dur
 (tur??? kelmeaktur,

yoqtur)
present-future
tense
 GY=du
neutral enclitic
 GZ=di
suppositional
past
 HA=tu
 HB=ti
neutral
subjective
ending
 HC=ken
neutral hearsay
 HD=imish
 HE=mish
dubitative
 HF=ghu
 HG=qu
 ??? = du, tu
vowel-neutral
marker
 HH=dek
vowel-neutral
adverb-forming
suffix ???
 HI=che
interrogative
 HJ=mu
Nonamed ???

HK=men
 HL=sen
 HM=siz
 HN=la
 HO=le
 HP=li
 HQ=miz
 HR=siler
 HS=sizler
interrogative
 HT=mu
 HU=mikin
dubitative
 HV=ghu
 HW=qu
hearsay
 HX=imish
 HY=mish
limitative
particle
 HZ=la
final (post
« mish »)
interrogative
 IA=mu
exclamative
final
 IB=e
 IC=a

Table des alphabets de langues turques

			Азери	Казакх	Киргиз	Татар	Turc	Turkmén		Оуighour						
			Азери	Казакх	Киргиз	Татар	Turc	Turkmén		Оуighour						
L	Cy	Ar	Азери	Казакх	Киргиз	Татар	Turc	Turkmén		Оуighour						
a			а	а	а	а	а	а		а						
'	'		я	я	я	я	я	я		я						
a	а	ا	а	а	а	а	а	а		а					zh	
b	б	ب	б	б	б	б	б	б		б					ch	
c	ч	ج					ж	ж		ж					sh	
ç	ч	چ	ч	ч	ч	ч	ç	ç		ч				ch	q	c
d	д	د	д	д	д	д	d	d		д				d	d	d
e	е		е	е	е	е	e	e		е				e	e	e
f	ф	ف	ф	ф	ф	ф	f	f		ф				f	f	f
g	г	گ	г	г	г	г	g	g		г				g	g	g
ğ	Ғ	غ	Ғ	ع		ع	ğ	Ғ		غ				gh	q	g
h	h	ح	h	ه			h	h		ه				h	h	h
i	и	ی	и	ي	и	ی	i	и		ی				i	i	i
ı	ы		ы	ی	ы	ی	ı	ы		ы						
j	ж	ژ	ж	ج	ж	ج	j	ж		ژ			j,zh	z		
k	k	ک	k	ك	k	ك	k	k		ك			k	k	k	k
l	л	ل	л	ل	л	ل	l	л		ل			l	l	l	l
m	м	م	м	م	м	م	m	м		م			m	m	m	m
n	н	ن	н	ن	н	ن	n	н		ن			n	n	n	n
o	о	و	о	و	о	و	o	о		و			o	o	o	o
ö	ө		ө	و	ө	و	ö	ө		و			ö	ө	ө	ө
p	п	پ	п	پ	п	پ	p	п		پ			p	p	p	p
q	г	ق	қ	ق		ق	q			ق			q	қ	қ	қ
r	р	ر	p	ر	p	ر	r	p		ر			p	r	r	r
s	c	س	c	س	c	س	s	c		س			s	c,s	s	s
ş	ш	ش	ш	ش	ш	ش	ş	ш		ش			sh	x	s	s
t	т	ت	т	ت	т	ت	t	т		ت			t	t	t	t
u	y		y	ۇ	y	ۇ	u	y		ۇ			u	u	u	u
ü	ү		ү	ۇ	ү	ۇ	ü	ү		ۇ			ü	ü	ü	ü
v	в		в	و	в	ۇ	v,w	в		و			w	w,v	w	w
x	x	خ	x	ح	x	ح	x	x		خ			x	h	x	x
y	j	ی	й		й	ي	y	й		ي			y	y	y	y
z	з	ز	з	ز	з	ز	z	з		ز			z	z	z	z
ə	ə	ع	ə	ا			ə	ə		ا			e	ə	a	a
			ë		ë		ë			ë			ë			
			i	ی			í									
		گ	ң	ڭ	ң	ڭ	ñ	ң		ڭ			ng	ng	n	n
			ұ	ۇ												

Notes : 0) Cette table permet de présenter de manière synthétique les prononciations similaires de lettres, et de comparer des caractères utilisés par les 8 différentes langues turques. 1) L'ordre alphabétique est soumis au tri informatique de l'azéri. Il ne représente pas l'ordre alphabétique des langues existantes. 2) L'ordre de priorité des alphabets de chaque langue est : l'écriture officielle, puis le nombre d'utilisateur. 3) Abréviations des alphabets: Ar – arabe, Cy – Cyrillique, La – latin, Pinyin et UKY : voir 2.2.1.

Abréviations utilisées

A => B	A results in B
A:B	B is the realisation at the surface level of A (the lexical level)
*A	is preceded by another letter
A*	is followed by another letter
A*B	A is followed by B
A__B	A is the left context of B and B is the right context of A
→	becomes
C	consonant
C _h	unvoiced consonant (hard consonant)
C _s	voiced consonant (soft consonant)
D	lexical dental consonant realized as {d} or {t} at the surface level
K	lexical laryngo-velar consonant realized at the surface level as {g} {k} {gh} or {q}
V	vowel
V _f	front vowel
V _b	back vowel
V _r	rounded vowel
V _u	un-rounded (high) vowel
A	high vowel {a} {e}
A _w	weakened vowel
I	high vowel {i} {ü} {u}
ms	monosyllabic
Rad	radical
Suf	suffix
Dub	dubitative

Code	Signification
1,2,3	1 ^{er} , 2 ^{ème} , 3 ^{ème} personne
ABL	Ablative
ABS	Abstract
ACC	Accusative
ADJ	Adjective
ADV	Adverbe
DAT	Dative
DET	Déterminant
DIR	Directif
F	Féminin
FRM.	Formel
GEN	Génitif

IFRM	Informel
INT	Interrogatif
LOC	Locative
M	Masculin
N	Nom
PAST	Passé Composé
PL	Pluriel
POS	Possessif
PREF	Préfixe
PREP	Préposition
PRO	Pronom
REL	Relatif
S	Sujet
SG	Singulier
SUFF	Suffixe
SuffDer	Suffixe dérivatif
T	Temps
V	Verbe
VM1	Impératif
VTP1	Imparfait

Groups des suffixes verbaux

Name	Abr.	Suffixes	future	FUT	i, y
euphonic vowels	EuV	i ,u,ü	future participle	PFU	dighan, dighin
euphonic consonant	EuC	y, s	positive		
reflexive	REF	n	speculative	SFP	ar, er, r
first passive	PA1	l	future negative		
second passive	PA2	l	speculative	SFN	s
aspectual auxiliary	VAA	wet, wét, wer, wér, wal, wél, wat	future conditional	CND	sa, se, si
first collective	CO1	sh	deverbal n. / adj. former	DVN	chi, ch
second collective	CO2	sh	deverbal substantive	DVS	lik, liq, luq, lük
factive	FAC	dür,tür, dur, tur	former		
potential	POT	ele, eli, el, ala, ali, al	privative	PRV	siz
negative	NEG	me, ma, mi	Pre-past tense		
repetitive	REP	e, a	interrogative	INP	m
connective	CON	p, y, w	particle		
purpositive	PUR	gili, kili, ghili, qili	vowel neutral auxiliary	VAT	i
simultaneative	SIM	gech, kech, ghach, qach	perfective past	TPP	di, ti, du, tu, dü, tü
interrogative	INT	em, am, m			m, ng, ngiz, i, si,
infinitive	INF	mek, mik, maq, miq			miz, k, q, nglar, liri,
substantiative	SUB	sh	possessive	POS	im, ing, ingiz, imiz, inglar
desiderative	DES	ghu, qu, gü, kü, ghay, qay, gey, key	relative	RLC	diki, tiki
participial past	PPA	ghan, qan, gen, ken, gin, kin, ghin, qin	locative		
adverbiative	ADV	séri	plural	PLR	lar, ler, lir
polite imperative	IM1	ghay, qay, gey, key	locative	CLC	de, te, di, ti, da, ta
imperative	IMP	sun, ey, ay, y	dative	CDA	gha, qa, ge, ke, ghi, qi, gi, ki,
habitual past	PHA	et, at, yt	ablative	CAB	din, tin
			accusative	CAC	ni

genitive	CGN	ning	adverb of manner	AMN	che
relative genitive	CGP	ningki	limitative particle	LIM	la
declarative	DEC	dur, tur	Interrogative 1	IN1	mu
present-future tense	TPF	du	Interrogative 2	IN2	mu, mikin
vowel neutral enclitic	EDI	di	Interrogative 3	IN3	mish
suppositional past enclitic	ETU	tu	exclamative final	EXC	a, e
suppositional past enclitic	ETI	ti	second set personal pronouns	PP2	men, sen, siz, la, le, li, miz, siler, sizler
subjective enclitic	ESU	ken			
hearsay enclitic	EHS	mish, imish			
dubitative	DUB	ghu, qu, du, tu			
equative	EQU	dek			

Lettres ouïghoures et leurs valeurs Unicode conventionnées²²⁵

ئاساسى رايۇن (06FF-0600)	يالغۇز	باش	ئوتتۇرا	ئاياق
FE8E ا				FE8D ا
FEEA ه				FEE9 ه
FE90 ب	FE92 ب	FE91 ب	FE8F ب	0628 ب
FB57 پ	FB59 پ	FB58 پ	FB56 پ	067E پ
FE96 ت	FE98 ت	FE97 ت	FE95 ت	062A ت
FE9E ج	FEA0 ج	FE9F ج	FE9D ج	062C ج
FB7B چ	FB7D چ	FB7C چ	FB7A چ	0686 چ
FEA6 خ	FEA8 خ	FEA7 خ	FEA5 خ	062E خ
FEAA د			FEA9 د	062F د
FEAE ر			FEAD ر	0631 ر
FEB0 ز			FEAF ز	0632 ز
FB8B ژ			FB8A ژ	0698 ژ
FEB2 س	FEB4 س	FEB3 س	FEB1 س	0633 س
FEB6 ش	FEB8 ش	FEB7 ش	FEB5 ش	0634 ش
FECE غ	FED0 غ	FECF غ	FECD غ	063A غ

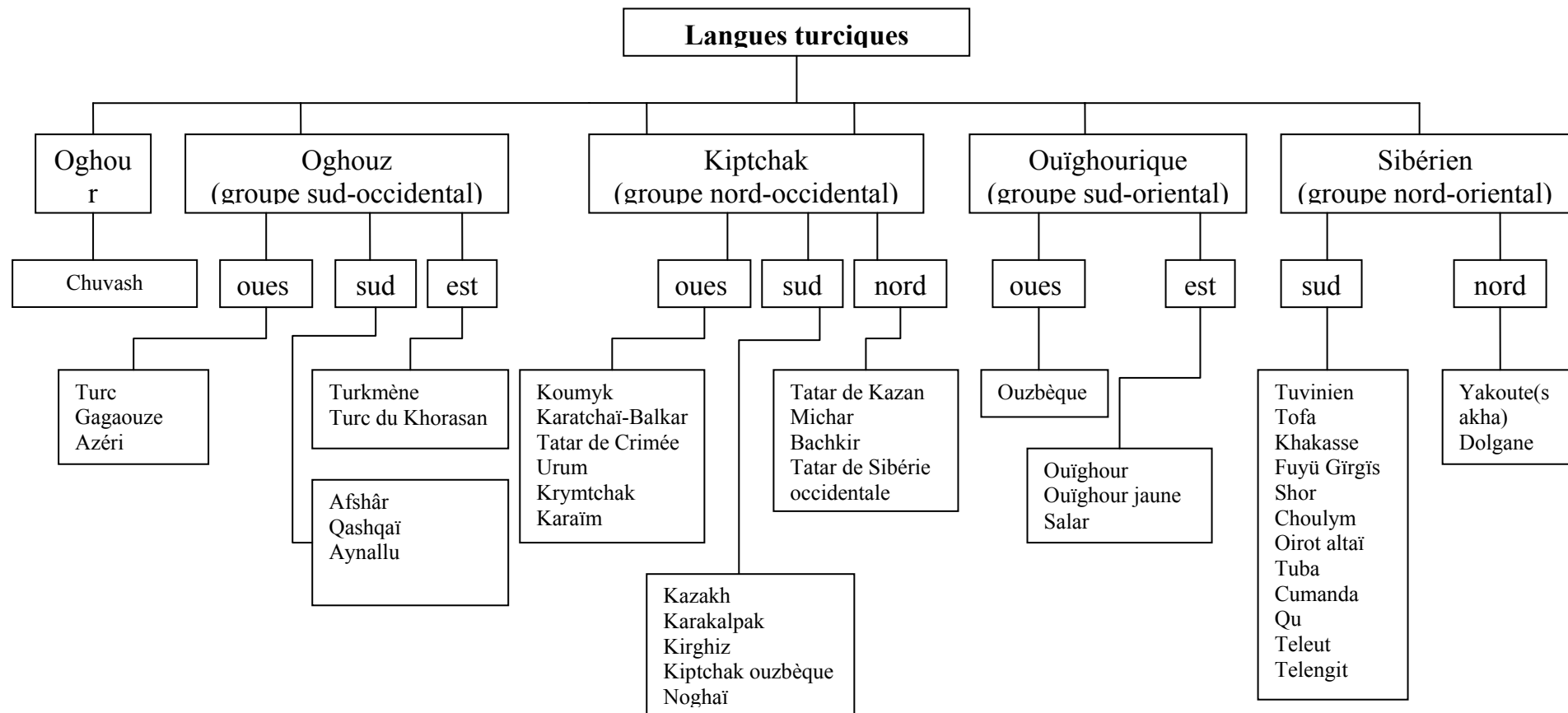
²²⁵ le 17 décembre 2004 par l'UCSA, voir <http://ukij.org/fonts>

FED2 ف	FED4 ف	FED3 ف	FED1 ف	0641 ف
FED6 ق	FED8 ق	FED7 ق	FED5 ق	0642 ق
FEDA ك	FEDC ك	FEDB ك	FED9 ك	0643 ك
FB93 گ	FB95 گ	FB94 گ	FB92 گ	06AF گ
FBD4 ڭ	FBD6 ڭ	FBD5 ڭ	FBD3 ڭ	06AD ڭ
FEDE ل	FEE0 ل	FEDF ل	FEDD ل	0644 ل
FEE2 م	FEE4 م	FEE3 م	FEE1 م	0645 م
FEE6 ن	FEE8 ن	FEE7 ن	FEE5 ن	0646 ن
FBAB ه	FBAD ه	FBAC ه	FBAA ه	06BE ه
EEEE و			FEED و	0648 و
FBD8 ۇ			FBD7 ۇ	06C7 ۇ
FBDA ۆ			FBD9 ۆ	06C6 ۆ
FBDC ۈ			FBDB ۈ	06C8 ۈ
FBDF ۋ			FBDE ۋ	06CB ۋ
FBE5 ې	FBE7 ې	FBE6 ې	FBE4 ې	06D0 ې
FEF0 ى	FBE9 ى	FBE8 ى	FEF ى	0649 ى
FEF2 ي	FEF4 ي	FEF3 ي	FEF1 ي	064A ي

FE8C ئ	FE8B ڭ	FEFC لا	FEFB لا
-----------	-----------	------------	------------

(FE8C)«ڭ»، (FE8B)«ئ» لارنى كىرگۈزۈشتە ئاساسى رايۇندىن (0626)«ئ» ئىشلىتىلىدۇ			
0648+0626 (ئ + و) ئو	0627+0626 (ئ + ا) ئا	0627+0644 (ا + ل) لا	...

Arbre des langues turciques



Les langues turciques: une trentaine de langues issues du turc ancien et encore parlées par plus de deux cent millions de locuteurs naturels et divisées en cinq familles principales: Oghour, Oghouz, Kiptchak, Ouïghourique et Sibérien. Après les cinq familles, toutes ces langues partagent les mêmes caractéristiques d'agglutination et, à l'exception de l'ouzbèque, d'harmonie vocalique. Cet arbre est inspiré de la classification présentée par Lars Johanson (1998), voir : http://en.wikipedia.org/wiki/Turkic_languages.

Liste des verbes primitives

ئاپار	ئالچاي	ئىگىلە	ئۇسسۇ	ئېلىشتۇر	بوغ
ئات	ئالدا	ئىل	ئۇق	ئېيت	بوغۇملا
ئاتا	ئالماش	ئىلغا	ئۇگىلا	ئەپلە	بول
ئاتلا	ئايا	ئىللىت	ئۇلا	ئەت	بويلا
ئاتلان	ئايرا	ئىغرا	ئۇنا	ئەچۈش	بىرىك
ئاجرا	ئايلا	ئۆت	ئۇيۇ	ئەدە	بىقسا
ئاچ	ئاينا	ئۆچ	ئۇۋا	ئەز	بىل
ئاچلا	ئاۋايلا	ئۆرتە	ئۇۋۇ	ئەزۋەيلە	بىلىجىرلا
ئاختۇر	ئاۋۇ	ئۆركە	ئۇۋۇلا	ئەسنە	بىلە
ئاداش	ئوتا	ئۆرلە	ئۈز	ئەش	بۆس
ئادالا	ئوخشا	ئۆرىدە	ئۈس	ئەكچە	بۆگ
ئارت	ئور	ئۆرۈ	ئۈسۈ	ئەكر	بۆل
ئارسىن	ئورا	ئۆزگەر	ئۈششۈ	ئەكىلە	بۆلە
ئارىلا	ئورنا	ئۆس	ئۈشقىر	ئەكە	بۇرا
ئاز	ئوزدۇر	ئۆكسۈ	ئۈشكۈر	ئەگ	بۇرۇق
ئاس	ئوسۇر	ئۆكۈن	ئۈگدە	ئەگىت	بۇرۇل
ئاسرا	ئۈشتۈ	ئۆگەن	ئۈلگۈر	ئەگەش	بۇز
ئاسغدا	ئوقۇ	ئۆل	ئۈلەش	ئەم	بۇلا
ئاش	ئولاش	ئۆلچە	ئۈمچەي	ئەۋەت	بۈك
ئاغ	ئوي	ئۇپرا	ئۈملۈل	باجا	بېزە
ئاغرا	ئويغان	ئۇت	ئېچا	بار	بېسىق
ئاق	ئوينا	ئۇچ	ئېرت	بازغا	بېكىت
ئاقسا	ئىچ	ئۇخلا	ئېرىت	باس	بېكىن
ئاقلا	ئىشلە	ئۇز	ئېرىن	باق	بېكە
ئال	ئىشلەپچىق	ئۇزا	ئېرە	بوپىلا	بېيىت
ئالاي	ئىشت	ئۇزار	ئېسەدە	بودا	بەپپەپلە
ئالجا	ئىشەن	ئۇس	ئېسەكە	بوشا	بەتلە

چۆرىدە	جۆنە	تۈگە	توۋلا	تاپشۇر	بەر
چۆرۈ	جۆيلۈ	تۈلە	تست	تارا	بەلدە
چۆكە	جۈدە	تۈنە	تىرش	تاراشلا	پات
چۆگىلە	جۈلگۈ	تۈڭكەل	تىرىل	تارت	پاخپاي
چۆلە	چاپ	تېتا	تىرە	تارقا	پارقىرا
چۆم	چاپلا	تېجە	تىز	تاشلا	پارلا
چۇلغا	چات	تېرا	تىزىملا	تاقا	پارىلدا
چۇۋۇ	چاتا	تېرىك	تىق	تاقاش	پالا
چۈچۈ	چاچ	تېلىق	تىك	تال	پانا
چۈش	چارلا	تېنە	تىل	تالا	پانلا
چۈشەن	چاق	تېپىل	تىلە	تاللا	پىچ
چۈمكە	چال	تېڭىرقا	تىن	تام	پىش
چۈمكەن	چان	تېۋىن	تىي	تان	پۇتا
چېقىش	چانا	تەپ	تىڭشا	تاي	پۇرا
چېكىن	چوقۇ	تەر	تۆكە	تاڭ	پۇركە
چېلىق	چولا	تەز	تۆگۈر	توختا	پۇش
چېنىق	چوڭاي	تەش	تۆلە	تورا	پۇيپۇلا
چەكە	چىچ	تەگ	تۆنە	توزۇ	پۈت
چەكلە	چىدا	تەيلە	تۇت	توس	پۈتكۈز
چەنلە	چىرىلدا	تەۋرە	تۇتاش	توشا	پۇر
خارىلدا	چىرە	جابدۇ	تۇر	توشقۇز	پۇركە
خالا	چىق	جاۋىلدا	تۇغ	توشۇ	پۇرمىلە
خورا	چىگ	جورا	تۇنجۇق	توقۇ	پۇرۇشتۇر
داجا	چىلا	جوۋا	تۇي	تول	پۈكە
داۋا	چىڭا	جىجا	تۈجۈپىلە	تولغا	پۈكلە
دورا	چۆچۈ	جىلماي	تۈر	تونۇ	پەتىلە
دومساي	چۆچۈت	جىما	تۈزە	توي	پەرلە
دومىلا	چۆرگىلە	جۈندە	تۈگ	توڭ	تاپ

قىچقار	قانا	شۈمشەي	سۈپۈر	سوي	دسلا
قىچىش	قاي	شۈمۈر	سۈر	سويۇ	دىگەلە
قىدىر	قايت	شەلۋەرە	سۈرۈشتۈر	سوۋۇ	دىۋىن
قىر	قايرا	غاجا	سۈز	سججا	دۈگجۈ
قىرقا	قايمۇق	غاڭشا	سۈمۈر	سىز	دۈگدەرە
قىرقىرا	قاينا	غودۇڭشا	سۈيلە	سىغ	دۈگدەي
قىرلا	قاڭتاي	غېمىسسا	سۇڭگۇ	سىغدا	دۈگىلە
قىزا	قاڭتۇر	غۇدۇرا	سىپسا	سىق	دېۋەيلە
قىزار	قاڭسا	غۇلا	سىپىغىن	سىك	دە
قىزىت	قاڭقا	قات	سىپلىشتۇر	سىلا	دەڭسە
قىس	قاڭقىرا	قاتنا	سىپلىك	سىلكە	رۇۋۇكلا
قىستا	قاۋا	قاتناش	سەپ	سىنا	زىبىلدا
قىل	قوپ	قاچ	سەرە	سىي	زەپلە
قىنا	قوپۇر	قاچىلا	سەز	سىيپا	ساپ
قىي	قوچۇ	قادا	سەگە	سىيرا	سات
قىيساي	قوداڭشا	قارا	سەيلە	سىيلا	ساغ
قىيغات	قورشا	قارغا	سەڭكە	سىك	ساقلا
قىينا	قورق	قاز	شاللا	سۆرە	سال
قىغىغاي	قورۇ	قاسا	شايللا	سۆزلە	سانا
قۇتۇل	قوزغا	قاغىجىرا	شورا	سۆك	ساۋا
قۇچ	قوش	قاق	شىل	سۆي	سورا
قۇر	قوغدا	قال	شىلا	سۇغۇر	سورۇ
قۇرۇ	قومۇر	قالا	شىلتا	سۇق	سوز
قۇس	قون	قالدۇر	شىنتاي	سۇقۇن	سوغتان
قۇي	قوي	قاما	شۇڭغۇ	سۇلتاي	سوغۇل
قۇيىقلا	قوۋۇ	قامدا	شۈركۈن	سۇن	سوق
قېرىش	قىتلا	قاملاش	شۈم	سۇيۇل	سولا
قېيدا	قىتىڭشا	قان	شۈمدەكلە	سۇۋا	سولاش

يېپىش	يىرگەن	ياق	مىلە	كەر	كالچاي
يېتىرغا	يىغ	يالا	مىن	كەز	كاي
يېتىل	يىغلا	يالتىرا	مىنگەش	كەس	كوچىلا
يېتىلە	يىغىل	يالا	مۆرە	كەل	كولا
يېلىن	يىقىت	يالىما	مۆك	كەملە	كىر
يېپىت	يىقىل	يالىلدا	مۇجۇ	كەي	كىشەنە
يېپىل	يىمىر	يالۋۇر	مۇقۇ	گۈلەي	كۆچ
يە	يۆتكە	ياما	مۇخداش	گۈگۈكلا	كۆر
يەت	يۆتەل	يان	مۈك	گېزىر	كۆرسەت
يەش	يۆگە	يانجا	مۈكچەي	گەۋدىلە	كۆشە
يەملە	يۆلە	ياي	مۈگدە	لاۋۇلدا	كۆكە
يەڭ	يۆمە	ياغرا	مۈنۈرە	لوغۇلدا	كۆم
ھار	يۆنەل	يوپۇت	مەرە	لېيا	كۆمتۈر
ھودۇق	يۇ	يوپۇر	نوقۇ	لەيلە	كۆن
ھولۇق	يۇت	يودا	نۆتۈرە	ماتا	كۆي
ھوماي	يۇغۇر	يورغىلا	ياپ	ماختا	كۆيدۈر
ھىجارا	يۇق	يورۇ	يات	مارا	كۈت
ھىجاي	يۇل	يوشۇر	يار	ماڭ	كۈچە
ھىلىپىرلا	يۇم	يوقا	يارا	ماڭدا	كۈرسۈ
ھېيىق	يۇمشا	يولا	يارمىدا	موتا	كۈل
ۋايسا	يۇمىلا	يولۇق	ياز	مولا	كېكىر
	يۇدۇ	يۈنۇ	ياسا	مىجا	كېڭەي
	يۇر	يىر	ياشا	مىرقىرا	كەپلە
	يۈگۈر	يىرت	ياغ	مىقىرا	كەت